RiCL

Research in Corpus Linguistics

RiCL 9/1 (2021)

Special Issue

"Challenges in combining structured and unstructured data in corpus development"

edited by Tanja Säily and Jukka Tyrkkö

aelinco
Asociación Española de Lingüística de Corpus

RiCL 9/1 (2021)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

9/1 (2021)

Articles — *Pages*

Book Reviews

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

RiCL Research in Corpus Linguistics

Generating linguistically relevant metadata for the *Royal Society Corpus*

Katrin Menzel – Jörg Knappen – Elke Teich
University of Saarland / Germany

Abstract – This paper provides an overview of metadata generation and management for the *Royal Society Corpus* (RSC), aiming to encourage discussion about the specific challenges in building substantial diachronic corpora intended to be used for linguistic and humanistic analysis. We discuss the motivations and goals of building the corpus, describe its composition and present the types of metadata it contains. Specifically, we tackle two challenges: first, integration of original metadata from the data providers (JSTOR and the *Royal Society*); second, derivation of additional linguistically relevant metadata regarding text structure and situational context (register).

Keywords – corpus building and extension; specialized diachronic corpora; written scientific English discourse; *Royal Society Corpus*; register-based metadata

1. INTRODUCTION[1]

This paper provides an overview of metadata generation and technical metadata management solutions for the *Royal Society Corpus* (RSC). The RSC is a diachronic, specialized corpus of scientific English covering more than 330 years of scientific journal articles (1665–1996) with the majority of its texts representing Present-day English and a smaller part representing Late Modern English. The corpus has been built to examine the development of scientific English, that is, the linguistic reaction to specialization and diversification in the scientific domain in terms of style and register/sublanguage formation. Various corpus extensions with more textual and contextual data across several releases have enriched the original corpus version over the years so that the newest releases, RSC 6.0 Open and RSC 6.0 Full (Fischer *et al*. 2020), cover optimized OCR

[1] The work reported in this paper has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102 "Information Density and Linguistic Encoding" as well as the Federal Ministry of Education and Research (BMBF) as part of the German Common Language Resources and Technology Infrastructure (CLARIN-D). We are especially indebted to the Royal Society of London and Dr Louisiane Ferlier for making available the source data.

results, more fine-grained linguistic and metadata annotations and a considerably larger number of texts from a much longer time span than previous corpus versions (Kermes *et al*. 2016).

We address two challenges regarding metadata: integration of descriptive metadata from heterogeneous sources and derivation of additional, linguistically relevant metadata from the corpus texts themselves. We first provide an overview of the corpus and the goals and motivation for building it and present the most important metadata requirements (Section 2). We then show which types of metadata have been gathered, distinguishing between descriptive, structural and derived metadata, how they are represented and stored, how they have been checked for completeness, consistency and quality and what types of corrections have been made when deviations were observed (Section 3). Finally, we provide information on the availability of the RSC (Section 4). Section 5 concludes the paper with a brief summary.

2. Overview of the RSC: Corpus material, basic processing and desiderata for metadata

The RSC is a diachronic specialized corpus of scientific English covering more than 330 years of scientific journal articles (1665 to 1996). The primary motivation for building the corpus was to provide a resource for empirically investigating the diachronic development of scientific English (see Halliday and Martin 1993) and its subregisters (sublanguages of chemistry, physics, biology etc.). Another important goal was to create a fairly coherent, homogeneous resource for exploring to what extent the temporal dynamics of language is shaped by communicative concerns, such as efficiency, informativeness, (non-)redundancy and unambiguousness. In particular, we are exploring whether information density (Crocker *et al*. 2016) is an independent factor in language change or whether it correlates with specific extra-linguistic variables, for example, scientific vs. non-scientific domain of discourse (Degaetano-Ortlieb and Teich 2019).

The RSC is embedded in an ecosystem of corpora of English scientific texts such as the *Coruña Corpus of English Scientific Writing* (Moskowich 2012; Moskowich *et al*. 2019), the corpus of *Middle English Medical Texts* (Taavitsainen *et al*. 2005) and its companions for Early and Late Modern English (Taavitsainen and Pahta 2010; Taavitsainen and Hiltunen 2019), or *SciTeX* (Degaetano-Ortlieb *et al*. 2013), a diachronic

corpus of modern English scientific texts. For a discussion and comparison of these corpora to the RSC see Fischer *et al*. (2020).

Going beyond these specific interests, from the beginning, the RSC was built as a resource to be shared by a larger community. As a domain-specific corpus with nearly all full texts from selected prestigious scientific journals that have impacted science across the globe, the RSC is a unique resource for historical linguists and sociolinguists as well as historians of science. As two of the world's longest-running academic journals, the *Philosophical Transactions* and the *Proceedings of the Royal Society of London* used to cover all known scientific disciplines of the time. They split into more specialized series for specific disciplines as the breadth and scope of scientific discovery increased by the end of the nineteenth century to cover mathematical and physical sciences and biological sciences separately. Texts from a few other *Royal Society* journals from the twentieth century, such as *Notes and Records of the Royal Society*, covering the history of science and the history of the *Royal Society* as a scientific community, and the *Biographical Memoirs of Fellows of the Royal Society*, with biographical essays, are also part of the corpus. These can also be queried separately.

The kinds of linguistic studies enabled by the RSC include the diachronic study of selected constructions as pursued, for example, in Construction Grammar, lexical-semantic change, sociolinguistic change, diachronic terminology development and register studies looking at language use according to situational context (field / topic, mode / medium and tenor / attitude of discourse). Metadata on discourse fields, for instance, enable the comparison of different scientific disciplines and help to reveal interesting differences in diachronic developments across disciplines (Teich *et al*. 2016).

2.1. Basic corpus data and processing

The first version of the RSC (2.0) was compiled for the time period of 1665–1869 (ca. 32 million tokens) on the basis of data obtained from JSTOR[2] (Kermes *et al*. 2016) and subsequently enlarged with texts from 1870 to 1996 obtained directly from the *Royal Society* (Fischer *et al*. 2020). We use metadata obtained from the *Royal Society* also for the texts obtained from JSTOR (see Section 3.3). With a size of around 48,000 texts and ca. 300 million tokens, the RSC now contains all English documents of the *Philosophical*

[2] http://www.jstor.org/

Transactions and *Proceedings of the Royal Society of London* and its more specialized successor journals from 1665 to 1996 (see Table 1).

Time period	Tokens
1665–1699	2,582,856
1700–1749	3,414,796
1750–1799	6,342,780
1800–1849	9,112,563
1850–1899	37,313,575
1900–1949	66,051,178
1949–1996	173,147,836

Table 1: *Royal Society Corpus* V5.1.0 (1665–1996)

After OCR optimization, normalization using VARD (Baron and Rayson 2008) was applied and all changes obtained by the normalization procedure were annotated into the corpus. We then added the standard linguistic annotations lemma and part of speech (UPenn tagset) automatically to all of our data using *TreeTagger* (Schmid 1994). In a final step, we added annotations for special research questions, including results of surprisal analysis (Knappen *et al*. 2017). One of the characteristics of electronic corpora is that text elements that are usually of minor importance for linguistic analysis and that are generally difficult to integrate or display correctly in linguistic corpora are typically removed during corpus building. This concerns particularly details of the layout, typographical markup, inserted material such as figures, tables or formulae, which are ignored and removed from the electronic text version. The same applies to elements of the page layout like headings and footers. We also removed hyphenation, even when the hyphenation crosses a page break. However, to also enable studies taking such elements into account, the RSC texts have been linked to their respective source texts on the *Royal Society* journal websites so that visual and layout elements in the image-based PDF files from the scans of the original documents can also be taken into account for individual analyses. The final product is an annotated corpus in the so-called vertical file format (.vrt, see Kermes *et al*. 2016) ready for import into the *Open Corpus Workbench* (Evert and Hardie 2011) and *CQPweb* (Hardie 2012). The .vrt format is a line-oriented file format with one token and all its annotations per line, interspersed by some simple XML-type markup lines. It is not a full XML format because of limitations in tag nesting and because of its line format.

2.2. Requirements on metadata

In accordance with the goals of providing a corpus for linguistic and humanistic study of scientific writing in Late Modern and Present-day English, from the outset, the metadata collected for the RSC provide as much information as possible about potentially relevant extra-linguistic variables. This clearly goes beyond the kinds of 'descriptive metadata' that typically come with datasets provided by digital archives, such as title, author, place etc. Additional metadata need to be derived from the texts themselves or by linking documents up with external sources, such as biographical databases of authors (see also Burnard 2005). Importantly, descriptive metadata and derived metadata have different functions for the user — descriptive metadata are necessary for 'identification' and 'discovery' (e.g. finding a relevant corpus through a data repository), derived metadata enhance the '(re)usability' of a corpus for an intended user community (e.g. facilitating the compilation of subcorpora according to discipline, time period, gender of authors etc.).[3] For the descriptive metadata coming from the text sources, we were faced with the additional challenge of the integration of two sets of metadata; as noted above, our sources came from two different archives (see Section 3.3 below). Two important steps with regard to derived metadata were to mark-up the logical text structure (e.g. title, abstract, text body), henceforth called 'structural metadata', which provides the possibility of integrating text structure elements as factors in analysis, and to assign discourse fields to the documents in the RSC which we realized using topic modelling, henceforth called 'contextual metadata'.

Other desiderata pertaining to formal aspects when a corpus resource is intended for use under FAIR principles are encoding standards (e.g. *Dublin Core*) and technical solutions, such as persistent metadata repositories (see Section 4).

3. RSC METADATA: TYPES OF METADATA, STANDARDS AND TECHNICAL SOLUTIONS

We start by contextualizing the issue of metadata in the context of the FAIR principles of data sharing and show our solutions (Section 3.1). Then we provide an account of the types of metadata we encode, distinguishing between descriptive and derived (structural

[3] See Section 3.1 below for more information on the FAIR principles of data sharing in relation to metadata.

and contextual) metadata (Section 3.2). Finally, we discuss the integration of metadata from heterogeneous sources (Section 3.3).

3.1. Realization of FAIR principles by metadata

The FAIR principles demand that a resource is Findable, Accessible, Interoperable and Reusable (see Table 2). Metadata are necessary for all four FAIR principles. Some of the FAIR principles, namely F4, A1 and A2, also address the necessity of a retrieval infrastructure. This infrastructure is described in Section 4.

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1. the protocol is open, free, and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be Reusable:
R1. (meta)data are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Table 2: The FAIR Guiding Principles (Wilkinson *et al*. 2016)

The metadata contain a persistent identifier for the corpus in a given corpus version; in our case, a handle from the *Handle System*.[4] The metadata describe the corpus in a rich way, allowing searches for corpora according to a variety of criteria. In this way, the FAIR principles F1–F3 (see Table 2) for Findability are fulfilled. The metadata also contain a description of the corpus, pointers to external resources like publications that describe the corpus and its building process in more detail and information on the copyright of the corpus. These metadata address the FAIR principle R1 (Reusability). The

[4] https://www.dona.net/handle-system

metadata for the whole corpus are provided in two formats, *Dublin Core*[5] and CMDI (Broeder *et al*. 2011). We follow the recommended vocabularies for *Dublin Core*, when applicable. The two formats, *Dublin Core* and CMDI, are standardized and highly interoperable, fulfilling the FAIR principles I1–I3 (Interoperability).

3.2. Types of metadata

3.2.1. Descriptive metadata

In terms of descriptive metadata, each document (text) includes a bibliographical identification of the text in traditional terms (author, journal, volume, pages, year of publication), as well as persistent identifiers to the sources (JSTOR IDs and DOIs from the *Royal Society of London*). This identification again relates to the FAIR principles F1–F3 (Findability) and R1.2 (Reusability) (see Section 3.1). The persistent identifiers enable the users of the corpus to go to photographic scans of the original text directly (see Figure 1 for an example).

Metadata for text *101322*	
Text identification code	101322
Journal in which the article was published	Philosophical Transactions (1665-1678)
Link to the source text on JSTOR	http://www.jstor.org/stable/101322

Figure 1: Excerpt from the metadata view in *CQPweb* showing a direct link to the JSTOR source

Descriptive metadata that provide classificatory information on the texts come from the JSTOR and *Royal Society* data. The *Royal Society* made a choice against relying on software which mines the data to extract titles, authors, dates, etc. and decided to employ indexers to manually catalogue the journals for various data. While we can extract and use these available data to complement our resource, some of them are more important to historians of science than for linguists.

The descriptive metadata are implemented in the .vrt file as attributes to the <text> tag that marks a single text in the corpus. We chose this way of encoding the textual metadata because it is compatible with the intended further processing of the corpus in the *Open Corpus Workbench* (Evert and Hardie 2011). For an overview of all descriptive metadata used in the RSC, see Table 3.

[5] https://dublincore.org/

Metadata type	JSTOR	*Royal Society*
author	✓	✓
title	✓	✓
journal	✓	✓
year	✓	✓
volume	✓	✓
first page	✓	✓
last page	✓	✓
issn	✓	✓
doi		✓
JSTOR id	✓	
language		✓

Table 3: Descriptive metadata taken directly from the sources

3.2.2. Structural metadata

We are concentrating on the text itself and we do not preserve most of its structural layout features, partly because they were not available in our sources (e.g. line breaks are not preserved in parts of the data and paragraphs are not marked), partly because non-linguistic elements like figures, tables or formulae are not directly relevant for linguistic study and are often badly represented in the OCR output. We also do not keep track of typographical markup like italicization. We remove recurring headlines and footers and keep only page breaks in the corpus. Pages are indicated by <page> tags and we add an attribute ID to this tag for the actual page number when we can get at it automatically and reliably. Pages are the only structural units still present in the processed text of the corpus, the titles are available as descriptive metadata to the texts and the abstracts or extracts are available as a separate corpus.

3.2.3. Contextual metadata

To approximate discourse fields, we computed topic models for various versions of the RSC (Fankhauser *et al*. 2016; Bizzoni *et al*. 2020). A topic model is a probability distribution over the words in the texts, and each text is composed from several topics. The topics are learned in an unsupervised fashion, but their labels are assigned manually by inspection of the most salient words. Figure 2 shows the hierarchical clustering of

topics for the RSC 6.0 Open. The five most characteristic words of each topic are given in the Appendix.

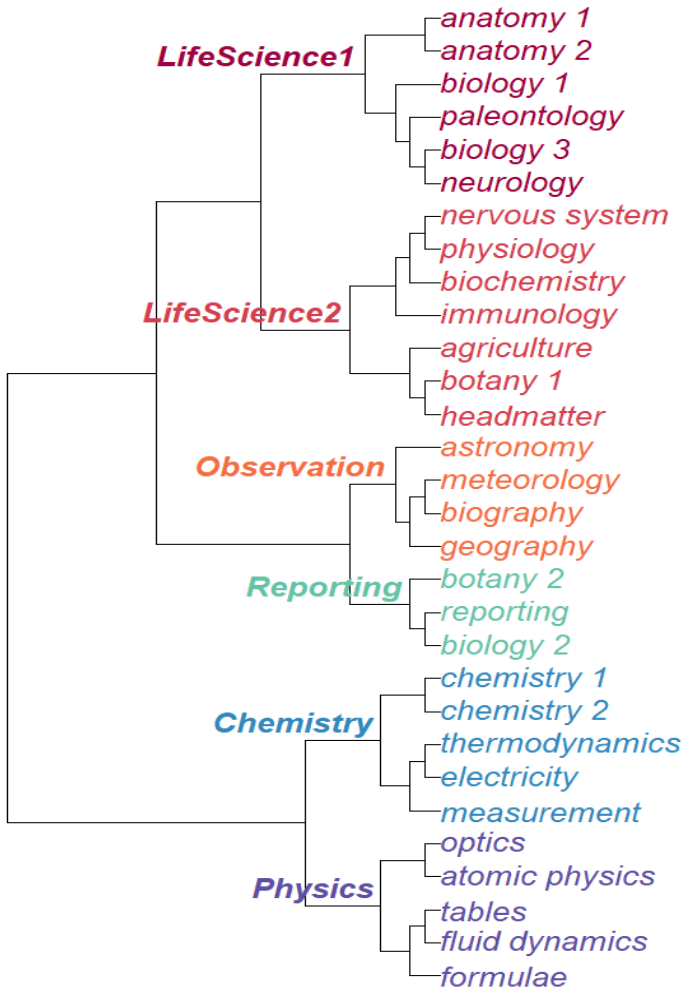

Figure 2: Topic hierarchy derived from topic modelling (RSC 6.0 Open)

As topic models provide not only word-topic but also document-topic assignments, we can add the topic labels as metadata to the documents contained in the corpus, as illustrated in example (1).

(1)
```
<text id="108995" issn="02610523" title="On Hydrofluoric Acid"
[…]
primaryTopic="Chemistry 2"
primaryTopicPercentage="74.1582515464929"
secondaryTopic="Thermodynamics"
secondaryTopicPercentage="12.760468963795098">
```

This is the basis for using topic information as an approximation of the fields of discourse of a text. We encode this information also as a CQP attribute, such that it can be used as a filter in corpus query.

3.3. Integration of metadata

In terms of identification, we set up a match between JSTOR IDs and RS DOIs based on basic bibliographic data: ISSN, volume, year, first page and last page. A match needs to be unique to be considered, as sometimes there are some different items on the same page. We did not use author and title information for this matching, as it decreases the recall significantly due to factors like differences in the encoding of special characters, such as apostrophes or accented letters. Not all articles from the JSTOR sources could be matched to DOIs. Apart from uniqueness there are also different factorings of the material into digital objects, like treating *An accompt on some books* either as single digital object or splitting it into several book reviews, or the treatment of errata and some coding errors. For those articles where DOI and JSTOR ID are matched for texts, the newly obtained metadata from the *Royal Society* are implemented also for previous corpus parts.

The other descriptive metadata types basically match across JSTOR and the *Royal Society* (RS) data. The main difference is that the RS dataset contains some additional and more specific information, such as markup of abstracts or extracts and article titles, contributor information (roles such as author, communicator, biographee or editor; affiliation, e.g. the university name; the *Royal Society* internal identifier number for RS fellows on the basis of which we can also gain further metadata on their biographical data, gender, etc.; election date to the RS), *MathML* markup of mathematical content as well as details on the publishing history.

Integration of the matching types of metadata (see Table 3) was straightforward. For the additional metadata included in the RS bundle, we pursued different strategies. For abstract/extracts (brief summaries of the corpus texts that were either available as abstracts of the respective texts or, in the absence of a given abstract, the first 200 words or the first paragraph of the body of the article), we decided not to add these texts to our corpus metadata but to use this information to create a separate additional corpus that only consists of the abstracts and extracts. Treating the abstracts as a corpus allows us to add linguistic annotations to the abstracts as well. In the case where the RS metadata were

more fine-grained than the ones from JSTOR, we made sure to retain as much detail as possible. For example, for ‘article-type’ for the first 200 years of the corpus we had used the categories ‘full article (fla)’, ‘book review (brv)’, ‘abstract (abs)’, and ‘obituaries (nws)’ in previous corpus versions where ‘fla’, ‘brv’ and ‘nws’ were taken directly from the JSTOR metadata and ‘abs’ was derived from the titles of the articles. For the RSC V6.0 we decided to use the finer grained text types from the *Royal Society* whenever a match was available and to drop the old text types from JSTOR. When no match was found we kept the JSTOR metadata. The vocabulary now includes: abstract, acknowledgement, addendum, appendix, article, astronomical, observation, bibliography, bill of mortality, biography, book review, catalogue, corrigenda, discussion, editorial, errata, experiment, index, lecture, letter, list, magnetical observation, meteorological observation, notes, obituary, preface, report, speech and symposium. Some of the text types like letter, speech or lecture give us a handle on the mode of discourse (e.g. written vs. written-to-be-spoken). For 10,397 texts where we have matched the metadata we see the following correspondence between the text types (Table 4).

We see a good match between the two systems, e.g. ‘brv’ (JSTOR) and ‘book-review’ (RS) are a very good match, and ‘abs’ (JSTOR) corresponds well with ‘abstract’ plus ‘paper-read’ in the RS data. The small category ‘nws’ from JSTOR containing obituary notes on deceased fellows is not represented as a separate article type but absorbed into the category of ‘article’ in the RS data. JSTOR’s ‘fla’ is divided into many subcategories. The majority of the texts, especially the later ones that have a much more standardized format, simply belong to the text type ‘article’. We deleted those texts from the corpus that only consist of tables or other non-text material (e.g. meteorological tables). The RS metadata also have a language attribute with a two-letter ISO 693 code (en, fr, es, la, it, sv, ro). We excluded those articles from the corpus whose main language is not English.

text type	abs	brv	fla	nws
abstract	2,060		560	
appendix			8	
article	35	27	3,421	5
astronomical-observation		1	434	
bill-of-mortality			8	
book-review		227	17	
catalogue			56	
editorial			3	
errata		1	9	
experiment		2	397	
illustration			1	
lecture			63	
letter		3	2,119	
list	1		1	
magnetical-observation			47	
meteorological-observation			134	
notes			3	
paper-read	16		2	
preface			4	
report	4		23	
speech			28	

Table 4: Correspondence between the *Royal Society* text types and our previous text type categories for the first 200 years of the RSC

Metadata concerning the authors of an article and their roles may also be of interest both to linguistic studies and to biographical studies on the authors. In Fischer *et al*. (2018) the authors were annotated and selected manually, matching different spellings of the name of the same person and separating authors with the same name because at that time no further author information was available. For the new release we include the fellowID received from the *Royal Society* whenever available and the author's role in the metadata for each text, as illustrated in example (2).

(2)
```
<text xmlns:xlink="http://www.w3.org/1999/xlink"
xmlns:mml="http://www.w3.org/1998/Math/MathML"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
id="rsta_1957_0024" issn="0080-4614"

title="The angular acceleration of liquid helium II"
fpage="359" lpage="385"

year="1957" volume="250" journal="Philosophical Transactions
of the Royal Society of London. Series A, Mathematical and
Physical Sciences"

author="H. E. Hall|D. Shoenberg, F. R. S."
fellowID="NA4060|NA5281"
```

```
authorRole="author|communicator" type="article"
corpusBuild="6.0"
doiLink="http://dx.doi.org/10.1098/rsta.1957.0024"
language="en">
```

From the *Royal Society* metadata on authors we use the fellowID (e.g. NA8137, named 'Code' in Table 5) uniquely identifying a fellow of the *Royal Society* and the authorRole when available. With the fellowID more biographical data of that specific author can be obtained. We have not added this additional information to the corpus yet as it needs some additional processing, but the fellowID is sufficient to link up to the information when needed. In the future, we intend to add nationality, gender (female first names of text authors or co-authors are often either spelled out or accompanied by the information *Miss*/*Mrs* in front of the initials of the first names) and the author's age (to be calculated from the birth date and year of publication if available).

Fellow details	
Surname	Boyle
Forenames	Robert
Epithet	Natural Philosopher and Chemist
Dates of Existence	1627 - 1691
Nationality	British
Dates and Places	Birth: Lismore Castle, Munster, Ireland (25 January 1627)
Address	Stalbridge Manor, Dorset (1645-1655); Oxford (1655-1668) Lady Ranelagh's house, Pall Mall, London (1668-1691)
Activity	Research Field: Natural philosophy, physics, chemistry
RS Activity	Membership: Founder Fellow Election Date: 28/11/1660 Council: Elected and declined Presidency of the Royal Society (1680)
Relationships	Fourteenth child, seventh son of Richard Boyle, 1st Earl of Cork, and his second wife, Catherine, daughter of Sir Geoffrey Fenton, Principal Secretary of State for Ireland [...]
Code	NA8137

Table 5: Example of Fellow details from the Royal Society Fellows Directory[6]

The author role helps us to identify who has actually written the article, who has communicated it, or who was taking part in a different role, for example, as an author of a reviewed book or as a biographee. This is useful to select works actually written by a

[6] https://royalsociety.org/fellows/fellows-directory/

certain author, for example, in order to determine the author's style or the development of an author over time. Many texts in the Late Modern English part of the RSC were submitted either by single individual authors who were Fellows of the Royal Society or by pairs of individual non-members and Fellows where the latter typically only acted as 'communicators'. Some prominent Fellows steered a large number of papers by non-Fellows through the publication process, often without having contributed to the actual research (cf. also Harrison 1989: 112). The proportion of multi-author papers has generally increased over time. Articles written by research teams become a common form in the Present-day English part of the RSC where it is not unusual to find research articles with four to ten authors, co-authors and other discourse participants.

4. Availability of the RSC

The corpus is deposited at a data repository at the certified CLARIN center of Saarland University.[7] CLARIN centers offer both direct web access to the metadata and an OAI-PMH interface for metadata harvesting. This guarantees that the corpus metadata are publicly accessible, addressing the FAIR principles A1 and A2 (accessibility). Large parts of the RSC have already been made available for free download and online query in a *CQPweb* interface from the CLARIN-D center at Saarland University under a persistent identifier.[8] Compared to the current release (V4.0), the next open version (V6.0 Open; Fischer *et al*. 2020) covers 50 additional years. Texts from certain decades currently remaining under copyright are not available for download as full texts, but the full version is available onsite. The CLARIN *Virtual Language Observatory* (VLO) harvests the metadata of the corpus and provides a facet search for corpora and language resources. The various elements in the CMDI metadata are mapped to the facets of the VLO and can be used to restrict search results (Van Uytvanck *et al*. 2012). This makes the RSC visible and fulfils the FAIR criterion F4.

5. Conclusion

We have shown how metadata contribute to the fulfilment of the FAIR principles and add value to a corpus for re-use by other researchers. We also note that metadata alone are not

[7] https://www.clarin.eu/content/clarin-centres
[8] http://hdl.handle.net/11858/00-246C-0000-0023-8D1C-0

enough to fulfil all FAIR principles: a retrieval infrastructure is also required. We used the *Royal Society Corpus* as a relevant example of how to obtain metadata, how to integrate them from different sources, and how to add some contextual metadata using topic modelling. For a summary of the metadata we discussed in this article see Table 6.

author	descriptive	doi	descriptive
first page	descriptive	last page	descriptive
title	descriptive	journal	descriptive
year	descriptive	volume	descriptive
issn	descriptive	JSTOR id	descriptive
language	descriptive	page	structural
primary topic	contextual	primary topic percentage	contextual
secondary topic	contextual	secondary topic percentage	contextual

Table 6: (Types of) metadata discussed in this article

The metadata provided for the RSC 6.0 Open allow for differentiated corpus analysis and query according to linguistically relevant variables such as time, author and topic (field of discourse) by selecting a subcorpus or comparing two or more subcorpora according to the metadata.

REFERENCES

Baron, Alistair and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics.* Birmingham, UK: Aston University. http://ucrel.lancs.ac.uk/people/paul/publications/BaronRaysonAston2008.pdf

Bizzoni, Yuri, Stefania Degaetano-Ortlieb, Peter Fankhauser and Elke Teich. 2020. Linguistic variation and change in 250 years of English scientific writing: A data-driven approach. *Frontiers in Artificial Intelligence – Language and computation, Research topic Computational Sociolinguistics* 3, Article 73.

Broeder, Daan, Oliver Schonefeld, Thorsten Trippel, Dieter Van Uytvanck and Andreas Witt. 2011. A pragmatic approach to XML interoperability – The Component Metadata Infrastructure (CMDI). In *Proceedings of Balisage: The Markup Conference 2011. Balisage Series on Markup Technologies 7.*

Burnard, Lou. 2005. Metadata for corpus work. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice.* Oxford: Oxbow Books, 30–46.

Crocker, Matthew W., Vera Demberg and Elke Teich. 2016. Information density and linguistic encoding (IDeaL). *Künstliche Intelligenz* 30: 77–81.

Degaetano-Ortlieb, Stefania, Hannah Kermes, Ekaterina Lapshinova-Koltunski and Elke Teich. 2013. SciTex: A diachronic corpus for analyzing the development of scientific registers. In Paul Bennett, Martin Durrell, Silke Scheible and Richard J. Whitt eds. *New Methods in Historical Corpora. Volume 3 of Korpuslinguistik und interdisziplinäre Perspektiven auf Sprache (CLIP).* Tübingen: Narr, 93–104.

Degaetano-Ortlieb, Stefania and Elke Teich. 2019[online]. Towards an optimal code for communication: The case of scientific English. *Corpus Linguistics and Linguistic Theory*. https://doi.org/10.1515/cllt-2018-0088

Evert, Stefan and Andrew Hardie. 2011. Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In *Proceedings of the Corpus Linguistics 2011 Conference*, Paper 153. Birmingham, UK: University of Birmingham. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2011/Paper-153.pdf

Fankhauser, Peter, Jörg Knappen and Elke Teich. 2016. Topical diversification over time in the *Royal Society Corpus*. In Maciej Eder and Jan Rybicki eds. *Digital Humanities 2016: Conference Abstracts*. Kraków, Poland: Alliance of Digital Humanities Organizations (ADHO), 496–500. https://dh2016.adho.org/abstracts/322

Fischer, Stefan, Jörg Knappen and Elke Teich. 2018. Using topic modelling to explore authors' research fields in a corpus of historical scientific English. In *Digital Humanities 2018: Book of Abstracts*. Mexico City, Mexico: Alliance of Digital Humanities Organizations (ADHO), 581–584. https://dh2018.adho.org/en/using-topic-modelling-to-explore-authors-research-fields-in-a-corpus-of-historical-scientific-english/

Fischer, Stefan, Jörg Knappen, Katrin Menzel and Elke Teich. 2020. The *Royal Society Corpus* 6.0: Providing 300+ years of scientific writing for humanistic study. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blace, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk and Stelios Piperidis eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 794–802. https://www.aclweb.org/anthology/2020.lrec-1.99.pdf

Halliday, Michael A.K. and James R. Martin eds. 1993. *Writing Science: Literacy and Discursive Power*. London: Falmer.

Hardie, Andrew. 2012. *CQPweb* – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17: 380–409.

Harrison, Andrew John. 1989. *Scientific Naturalists and the Government of the Royal Society 1850–1900.* The Open University, PhD dissertation.

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, Hélène Mazo, Asunción Moreno, Jan Odijk and Sterlios Piperidis eds. *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association, 1928–1931. https://www.aclweb.org/anthology/L16-1305.pdf

Knappen, Jörg, Stefan Fischer, Hannah Kermes, Elke Teich and Peter Fankhauser. 2017. The making of the *Royal Society Corpus*. In Gerolf Bouma and Yvonne Adesam eds. *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*. Gothenburg, Sweden: Linköping University Electronic Press, 7–11. https://www.aclweb.org/anthology/W17-0503.pdf

Moskowich, Isabel. 2012. CETA as a tool for the study of modern astronomy in English. In Isabel Moskowich and Begoña Crespo eds. *Astronomy "Playne and Simple": The Writing of Science between 1700 and 1900*. Amsterdam: John Benjamins, 35–56.

Moskowich, Isabel, Begoña Crespo, Luis Puente-Castelo and Leida Maria Monaco eds. 2019. *Writing History in Late Modern English – Explorations of the Coruña Corpus.* Amsterdam: John Benjamins.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Taavitsainen, Irma, Päivi Pahta and Martti Mäkinen eds. 2005. *Middle English Medical Texts*. Amsterdam: John Benjamins.

Taavitsainen, Irma and Päivi Pahta eds. 2010. *Early Modern English Medical Texts: Corpus Description and Studies*. Amsterdam: John Benjamins.

Taavitsainen, Irma and Turo Hiltunen eds. 2019. *Late Modern English Medical Texts: Writing Medicine in the Eighteenth Century*. Amsterdam: John Benjamins.

Teich, Elke, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes and Ekaterina Lapshinova-Koltunski. 2016. The linguistic construal of disciplinarity: A data mining approach using register features. *Journal of the Association for Information Science and Technology (JASIST)* 67/7: 1668–1678.

Van Uytvanck, Dieter, Herman Stehouwer and Lari Lampen. 2012. Semantic metadata mapping in practice: The Virtual Language Observatory. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asunción Moreno, Jan Odijk, Stelios Piperidis eds. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012).* Istanbul, Turkey: European Language Resources Association, 1029–1034. http://www.lrec-conf.org/proceedings/lrec2012/pdf/437_Paper.pdf

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

Corresponding author
Katrin Menzel
Department of Language Science and Technology
Saarland University, Saarbrücken Campus
66123 Saarbrücken
Germany
e-mail: k.menzel@mx.uni-saarland.de

received: February 2020
accepted: October 2020

APPENDIX

The following table gives the most characteristic words (word forms) for each of the thirty topics from the topic model in Section 3.2.3.

anatomy 1	fig plate cartilage part skull
anatomy 2	fig bone bones teeth surface
biology 1	number eggs species larvae female
paleontology	fig plate species form structure
biology 3	cells fig cell tissue nucleus
neurology	fibres posterior anterior fig side
nervous system	nerve muscle contraction stimulation muscles
physiology	blood serum action normal pressure
biochemistry	solution cent water acid vol
immunology	days growth water found bacteria
agriculture	nitrogen soil plants plot years
botany 1	fig plate section cells plants
headmatter	vol society london des der
astronomy	sun observations time stars distance
meteorology	observations days day p.m. magnetic
biography	society work years royal professor
geography	feet water sea found miles
botany 2	leaves plants plant fig species
reporting	great time made found account
biology 2	animal part blood parts body
chemistry 1	water air experiments quantity heat
chemistry 2	acid solution water obtained salt
thermodynamics	temperature pressure air gas tube
electricity	current wire resistance magnetic positive
measurement	inch fig inches made length
optics	light rays glass colour red
atomic physics	lines spectrum line bands spectra
tables	values table curve results case
fluid dynamics	velocity surface motion force direction
formulae	equation equations function form cos

RiCL Research in Corpus Linguistics

Corpus Linguistics and *Eighteenth Century Collections Online* (ECCO)

Mikko Tolonen[a] – Eetu Mäkelä[a] – Ali Ijaz[a] – Leo Lahti[b]
University of Helsinki[a] / Finland
University of Turku[b] / Finland

Abstract – *Eighteenth Century Collections Online* (ECCO) is the most comprehensive dataset available in machine-readable form for eighteenth-century printed texts. It plays a crucial role in studies of eighteenth-century language and it has vast potential for corpus linguistics. At the same time, it is an unbalanced corpus that poses a series of different problems. The aim of this paper is to offer a general overview of ECCO for corpus linguistics by analysing, for example, its publication countries and languages. We will also analyse the role of the substantial number of reprints and new editions in the data, discuss genres and the estimates of Optical Character Recognition (OCR) quality. Our conclusion is that whereas ECCO provides a valuable source for corpus linguistics, scholars need to pay attention to historical source criticism. We have highlighted key aspects that need to be taken into consideration when considering its possible uses.

Keywords – *Eighteenth Century Collections Online* (ECCO); *English Short-Title Catalogue* (ESTC); metadata; Optical Character Recognition (OCR); eighteenth-century studies; bibliographic data science

1. INTRODUCTION

The relevance of quantitative-statistical methods for the description of the variation of English has increased rapidly during the last decades (cf. Gries 2012). In sync with the increase of the relevance of statistical or quantitative approaches to language, the availability of real-time language data, instead of tightly controlled corpora, has become a feature of corpus linguistics (Davies 2012). For historical studies of language change, the availability of data is the key question as to the basis of any work in the field (Hiltunen *et al.* 2017). However, creating a representative corpus is often difficult. Informal spoken language rarely survives (see, however, Hitchcock and Shoemaker 2007), letter collections are highly selective (already because of the question of literacy

DOI 10.32714/ricl.09.01.03

rates) and printed documents are biased towards higher classes of language users. Most large digitised collections also come with precious little information on the balance and biases within the corpus.

In relation to the eighteenth century, *Eighteenth Century Collections Online* (ECCO) has recently received attention not only from historians but from corpus linguists as well.[1] For example, the *Linguistic DNA* project aimed to use it as one of the main sources to uncover 'the DNA' of historical English discourse.[2] There are good reasons to take ECCO as the basis of studies on language variation. It is the most comprehensive dataset available in machine-readable form for eighteenth-century printed texts. It is linked to the *English Short-Title Catalogue* (ESTC)[3] that enables linking the collection to complementary text sources structured in the same way (most importantly *Early English Books Online* (EEBO),[4] which contains publications from 1473 to 1700). At the same time, it poses a series of problems. In the *Linguistic DNA* project, it was quickly realised that the quality of Optical Character Recognition (OCR) is highly problematic. Their conclusion was that "there are too many problems within the OCR dataset to use it" (Linguistic DNA 2017). One community-driven solution to these problems has been the *Text Creation Partnership*, which has turned to manual work to produce accurate transcriptions of a portion of the titles for EEBO and ECCO.[5] However, whereas for EEBO the EEBO-TCP collection covers almost half of the EEBO texts, ECCO-TCP contains transcriptions for only 3,101 out of the more than 200,000 texts in total. Therefore, as the OCRed version of ECCO is a remarkable source in size and scale, it is important to continue efforts towards making use of it in a reliable manner (Bullard 2013).

A systematic large-scale analysis of the biases in large digitised collections, such as ECCO and ESTC, can be critically complemented by algorithmic approaches (Lahti *et al.* 2015; Tolonen *et al.* 2018; Lahti *et al.* 2019; Lathi *et al.* 2020; Tolonen *et al.* 2021). Data quality is often suboptimal, posing challenges for large-scale comparisons

[1] ECCO ids referenced can be queried through the web-interface at https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online
[2] https://www.linguisticdna.org/
[3] The ESTC ids referenced can be queried through the web-interface of the British National Library at http://estc.bl.uk, and all the information regarding individual records is accessible through it. ESTC records used in this article have been enriched from the state of the version behind the web-interface implementation.
[4] https://quod.lib.umich.edu/e/eebodemo/
[5] https://textcreationpartnership.org/

and research use. The need for large-scale harmonisation has been widely recognised, and various solutions that are relevant to corpus linguistics are already available or have been proposed for the processing of digitised texts and other data types (Mäkelä *et al.* 2020). Overall, the applications of data science in this context aim at systematic and scalable improvements in data harmonisation, enrichment, and analysis, with the ultimate goal of advancing research on digital resources. Our present work relies heavily on our earlier efforts to harmonise the ESTC bibliographic metadata and the ongoing work to assess and potentially improve the quality of the ECCO full text collection. Here, we take the first steps towards a systematic integration and joint analysis of these two complementary sources. Whereas statistical integration of data from heterogeneous sources is a topical area in contemporary machine learning research, many pragmatic issues related to data quality and biases need to be understood and overcome before systematic and reliable statistical analyses can be carried out.

According to Davies (2012: 172) the main problems with large text archives (such as ECCO) are "accuracy, annotation, architecture, availability, and genre balance between different time periods." In this paper, we will look particularly at availability, architecture, genre balance and the accuracy in terms of OCR quality. We weigh these aspects of ECCO and its use in corpus linguistics from different perspectives and especially with respect to selection of corpora. If the magnitude of ECCO as big humanities data is seen as its best asset, how comprehensive is it in fact? We have harmonised the ESTC and worked connecting ECCO to the ESTC so that we can, for the first time, statistically evaluate the range of ECCO in the light of the ESTC.[6]

The aim of this paper is to reflect on different aspects of ECCO, in particular from the perspective of corpus linguistics. In Section 2.1, we give a statistical overview of ECCO in terms of different countries where works in ECCO were published and languages used in ECCO. We will then turn to discuss the temporal distribution of ECCO over the eighteenth century. In Section 2.2, a crucial part of our analysis is the analysis of reprints and new editions in ECCO (Ijaz *et al.* 2019) and, in Section 2.3, we will discuss the subject topics and genres in ECCO. After this, in Section 2.4, we turn to discuss the OCR quality of ECCO before concluding our observations in Section 3.

[6] We are currently writing a separate comprehensive article about the representativeness in ECCO when compared to ESTC.

2. ANALYSIS

ECCO was released in 2002 as a web-based query platform, after which it has been widely used at different universities by researchers and students alike. Originally, ECCO was scanned in the late 1990s from microfilms that date as far back as the early 1980s. Later in the 2000s, Gale —the company that owns the rights to distribute ECCO outside Britain— launched ECCO Part II (ECCO 2) that added 50,000 titles to the collection. In total, there are currently over 200,000 titles in the collection. Gale is at present digitising more materials with the intention to launch ECCO Part III with approximately 90,000 new titles later. Thus, it needs to be understood that already by its basic makeup ECCO is not a carefully selected or let alone balanced collection, but a layered historical source (about the history and development of ECCO including the selection process, see especially Gregg 2020. See also Kinley 2003; Greenfield 2010; Gale 2016; Cayley 2017).[7]

The more than 200,000 eighteenth-century documents included in ECCO amount to a little over 50 per cent of what is included in ESTC, the most comprehensive metadata collection of the British publication record for the early modern period (1470–1800). Thus, when compared to the publication record in general, ECCO is an impressive collection. There are however clear imbalances in the collection. In this article, we will discuss particularly geographical distribution, languages, temporal distribution, genre and estimates of OCR quality.[8] All our calculations are based on XML data dumps of ECCO Parts I and II obtained from Gale in 2015 through the Helsinki University Library, in accordance with Gale's updated text mining policy that allows researchers of a subscribing institution access to the content outside of Gale's user interface. All comparisons to ESTC are against our offline version graciously provided to us by the British Library in March 2016 and updated later.

2.1. Place, language and dating of publications

If we look at the geographical distribution of works in ECCO (Table 1), we quickly realise that especially items printed in the US are heavily underrepresented in the collection, compared most importantly to Scotland and Ireland. This bias can mainly be

[7] We are very grateful to Stephen Gregg for sharing his monograph with us prior to publication.
[8] We are also working on an analysis of different authors in the collection, but it is beyond the scope of this article.

explained by the origin of the digitised documents in ECCO, where the main part originate from the British Library and, to an important degree, also Oxford and Cambridge. While American libraries have also been part of the projects underlying ECCO, it is still clear that they remain heavily underrepresented in the dataset.

Country	**ESTC**	**ECCO**
England	233,473	134,935 (58%)
Scotland	33,864	17,365 (51%)
Ireland	24,957	16,647 (67%)
USA	40,672	10,088 (25%)
France	2,527	1,398 (55%)
Canada	995	35 (4%)
Others	4,517	2,157 (48%)
Unknown	2,868	1,133 (40%)
Total	**343,873**	**183,758 (53%)**

Table 1: Countries of publication in ECCO and the ESTC[9]

English is, by a vast margin, the dominant language in the nationally built collections of ESTC and ECCO (cf. Table 2). It is partly a reflection of ongoing changes in the British society at the time, especially since the number of Latin works is remarkably low compared to, for example, the eighteenth-century German and French sources (Lahti *et al.* 2019: 15–17). The presence of Welsh materials is noticeable in ECCO, while particularly German sources are missing. Within English language publications, what is important for the study of language variation is that the number of publications in both Ireland and Scotland is high. Even when most of the Dublin printing activity focused on London reprints, there is still a good chance to use these materials to identify regional variation in language use in Britain. We consider this as one of the prominent research fields with respect to ECCO.

[9] The number of ESTC records for the same time period (1701–1800) is shown for comparison. The percentages indicate the fraction ESTC records that are coverd by ECCO. The aggregate 'Others' includes a mixed bag of all countries with fewer records in these collections than Canada, such as Belgium, Germany, Italy, Switzerland, and the Netherlands, but also Barbados, Haiti, India, Jamaica, and so forth. The Category 'Unknown' consists of records whose place of publication is recorded.

Primary language	ESTC	ECCO
English	324,804	173,967 (54%)
Latin	7,699	4,599 (60%)
French	7,269	3,783 (52%)
Welsh	765	540 (71%)
Italian	510	341 (67%)
German	1,630	279 (17%)
Others	1,196	249 (21%)
Total	**343,873**	**183,758 (53%)**

Table 2: Main languages in ECCO and the ESTC[10]

One aspect that needs to be taken into consideration when using ECCO for text mining is that the corpus is uneven over time. From 1780 to 1800 there are far more documents than during the earlier decades (cf. Figure 1). Since there are also changes in the distribution of genres during this time, this obviously is something that needs to be taken into consideration when using ECCO as a corpus.

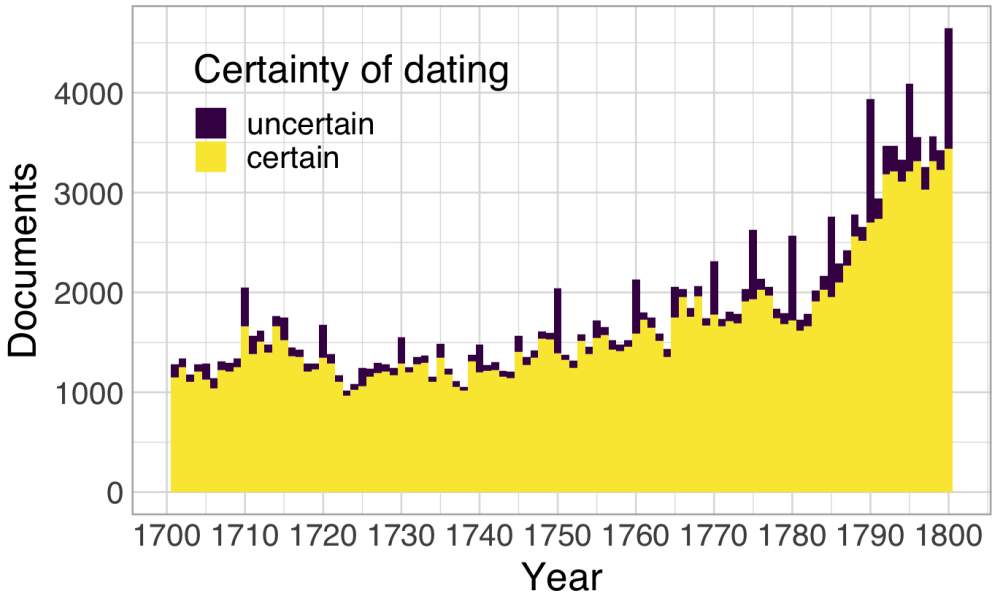

Figure 1: Variation in ECCO title count during the 18th century

Another important point is that some of the dates in ECCO are uncertain. Thus, a document dated for a particular year (particularly even years such as 1710) might actually be from any year during that decade. In many cases, the uncertainty has been indicated in the ESTC (with e.g. a question mark, 'ca.', or time range), and we have

[10] The aggregate "Others" includes Ancient Greek, Dutch, Hebrew, Portuguese, Spanish, Tamil, and a number of other languages.

used this to identify the uncertain years (shown in purple in Figure 1.). These uncertain attributions contribute peaks to even five, ten and 50 years.

2.2. Reprints

A further aspect that anyone using ECCO as a corpus needs to take into account is that a large part of the collection are reprints and further editions of previously published works. Gale, in their online materials, has suggested that new editions should contain substantial new material in order to be included, and that mere reprints would be for the most part excluded.[11] Based on our evaluation, however, this is not true and some titles are repeated dozens of times, years after their initial publication, while others are missing from the collection altogether.[12] This obviously has quite an impact on the general shifts in language that we might detect from the collection. One way of phrasing this is that we may get two different perspectives to language when using ECCO. If we use the collection as a whole, our perspective is the language available to readers at a particular time. Here it is obvious that if a particular work is printed verbatim several times over, it has more impact molding the minds of the reading public. We may look at the classics, for example, from this perspective. The other viewpoint would be to make a subset of ECCO that would include only any possibly novel parts of later editions past the first publication. If we are interested in neologisms, for example, this might be a more viable approach, because the dataset would only include new works and thus tracking the emergence and diffusion of new language might be easier.

With respect to duplication, two distinct viewpoints can be considered. First, we consider duplication within ECCO itself. Based on our analysis, in the 184,029 ESTC records contained in ECCO, there are 115,962 unique works. Therefore, a full 37 per cent of the content within ECCO may be duplicated elsewhere within it. Of the distinct works, 80 per cent appear inside ECCO only once, 11 per cent twice and nine per cent more than two times, with Thomas Sternhold and John Hopkins' *Book of Psalms* holding top place with 135 copies, followed by John Milton's *Paradise Lost* with 118.

---

[11] Originally, the *Eighteenth Century* microfilm project was limited to "first and significant editions of each title" with the exception of 28 major authors whose editions were all included (Alston 1981: 2). This is still visible in ECCO. For the full history of the complicated selection process behind ECCO, see Gregg (2020).

[12] For our process of identifying reprints, see Ijaz *et al.* (2019).

As a second viewpoint, we consider the amount of material in ECCO that are reprints from earlier years, without regard to whether the original versions are included in ECCO themselves. Via this viewpoint, we can consider how well the texts associated with a particular year in ECCO actually correspond to contemporary language, as opposed to the language of years past. As we notice in Figure 2, the fraction of material that are reprints from earlier years in ECCO grows particularly towards the 1750s up to >30 per cent.[13] In total, a full 31 per cent of the titles in ECCO are reprints of some kind, highlighting the increasing importance of reprints among the overall publishing activities. Naturally, new editions contain some new language (and, in some cases, also extensive additions to the original work), but eighteenth-century printing technology favoured exact reprinting (cf. Bonnell 2009). In terms of evaluating the bias caused by these reprints, it is interesting to know their age distribution. Here, the median age of the reprints is seventeen years from their first printing, but the spread is large, with a full nine per cent of reprints dating back more than 100 years ago.

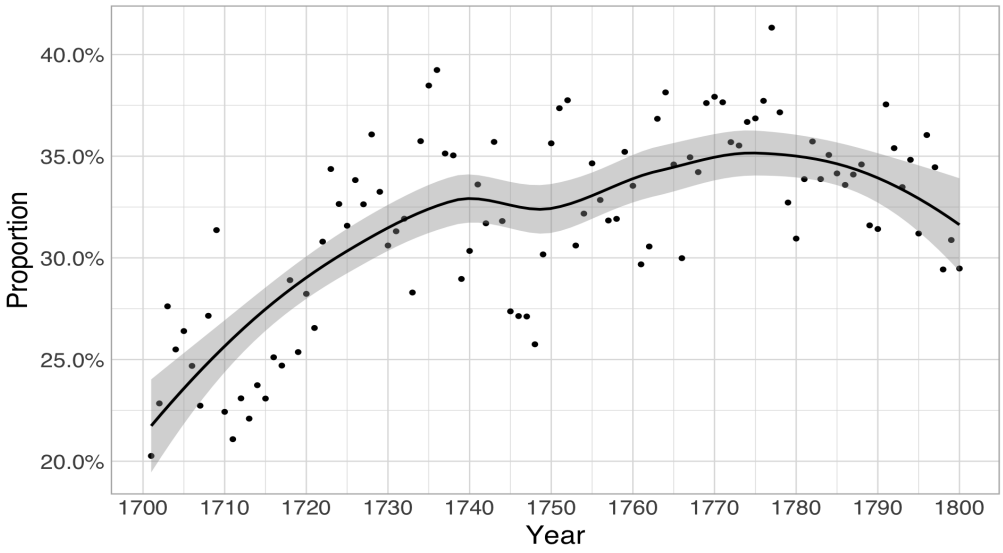

Figure 2: Share of ECCO that are reprints from earlier years for the period 1701–1800

More bias is added to this equation when we realise that the presence of popular authors in ECCO is prominent. This seems partly to be a legacy of the *Eighteenth Century* microfilm project where a decision was made to include all the editions of the works of

[13] The graph identifies the percentage of reprints each year as identified from ESTC. For our method of detecting reprints, see Ijaz *et al*. (2019). Also, texts printed before 1700 are included in the graph.

twenty eight authors considered 'major'.[14] As a result, the editions of, for example, Henry Fielding, Alexander Pope, Samuel Johnson and Laurence Sterne are nearly completely covered in ECCO, while the works of other eighteenth-century authors are not included, let alone all the editions of these works. This is a serious form of bias in the collection because it amplifies the effect of the already well-known and studied authors. Thus, we need to be careful not to take the language of Pope and Johnson, for instance, to represent the eighteenth century in general because of imbalances in the corpus that we study.

2.3. Subject headings

One feature of ECCO is that it includes subject headings for all the documents. This is also a legacy of the *Eighteenth Century* microfilm project where the collection was arranged to eight subject heading categories.[15]

When we examine the subject heading distribution over time, we realise that there are both lasting trends as well as spot anomalies in the data. Taking the proportion of running words in each section as a measure (cf. Figure 3 below), we see first of all that the share of 'Religion and Philosophy' goes down over the eighteenth century, whereas the role of 'Literature and Language' grows somewhat over time. At the same time, there is a significant anomaly in the 1730s where the proportion of words associated with 'General Reference' suddenly spikes upwards. Upon investigation, this spike is caused solely by the inclusion in ECCO of two separate 1734 editions of Pierre Bailey's dictionary, consisting of five and ten volumes of around a thousand pages each. Given that the language of such dictionaries is certainly a distinct genre with more precise definitions of words and concepts, not filtering these out may certainly affect any text mining results based on ECCO. Earlier we have examined this aspect with respect to use of philosophical language, and it turns out that towards the later

[14] Addison, Bentham, Bishop Berkeley, Boswell, Burke, Burns, Congreve, Defoe, Jonathan Edwards, Fielding, Franklin, Garrick, Gibbon, Goldsmith, Hume, Johnson, Paine, Pope, Reynolds, Richardson, Bolingbroke, Sheridan, Adam Smith, Smollett, Steele, Sterne, Swift and Wesley. For further discussion, see Gregg (2020).

[15] According to Gregg (2020: 21), "these subject headings may well have had their origin in Alston's experiments with the 18thC STC's initial online interface at the British Library, which he felt could help in the creation of subject packages which will form the basis of the RPI program to microfilm the substantive texts in ESTC (Alston 2004)."

eighteenth century the growth in precise definitions of philosophical concepts is considerable (Tolonen *et al.* 2017).

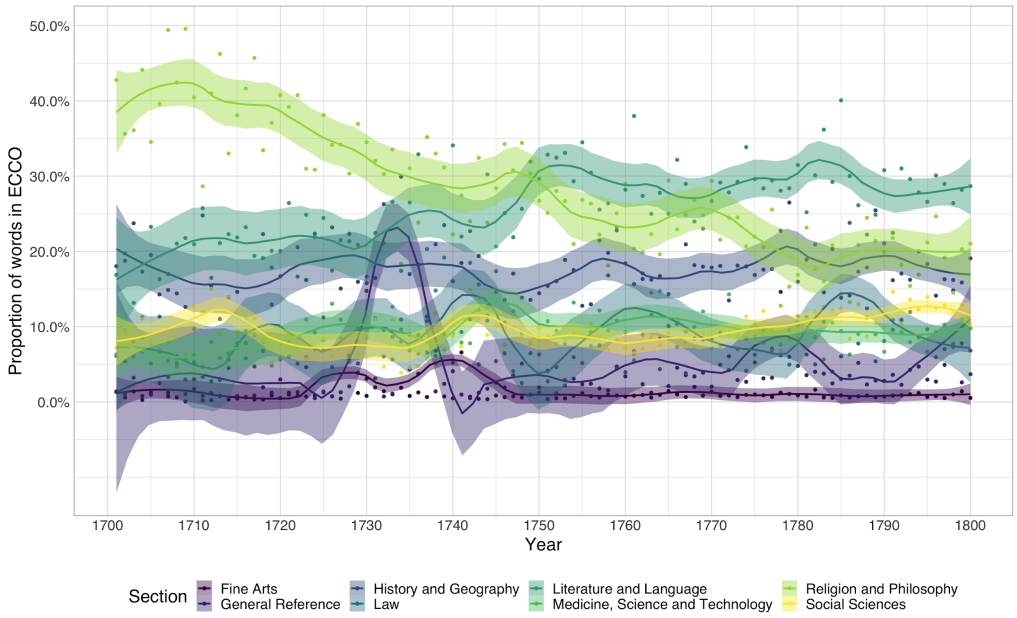

Figure 3: Composition of ECCO 1 in terms of the number of words by subject heading and year[16]

Apart from the few very large dictionaries and the anomalies they cause, the 'General Reference' and 'Law' categories, on the other hand, are much smaller in ECCO than they are in reality. This is because it is especially the almanacs, proclamations, general acts and the like that were intentionally excluded from the materials that form ECCO (Alston 1981). Yet also here there is a temporal anomaly. For reasons unknown to us, from the 1750s to the 1770s, a much larger amount of bills and petitions has been included. Due to these being very short, this anomaly is mostly not discernible in the 'Law' data of Figure 3 but does show up clearly if the data is weighted by the number of publications instead of the number of words in them.

2.4. OCR quality

As ECCO is a corpus arising from automated mass digitisation, it is susceptible to noise from the OCR process. In earlier work (cf. Hill and Hengchen 2019) comparing the ECCO-TCP hand-transcribed subset of ECCO 1 to the OCRed version, it was identified

[16] The integration of multi-volume titles and their impact on the numerical estimates are influenced by variations in publication years, edition counts, and other factors. A full manual curation of the large data collection is here replaced by an approximation, where the multiple volumes are aggregated and counted at the first occurrence. This makes it possible to scale up the estimates to cover the whole data collection but may introduce additional bias, such as the peak that we can observe at Bailey's 1734 dictionary.

that, on an overall level, the token-level mean precision of ECCO OCR is 0.744 (meaning that on average, 74% of the tokens in ECCO OCR are correct), with recall being 0.814 (meaning that 81% of the tokens in the original are included in the OCRed version).

While the above results speak directly only for the small ECCO-TCP subset, we also identified a statistically significant ($p<0.001$) Pearson correlation of 0.795 between the page-level F1 score (the harmonic mean of precision and recall) and the confidence value reported by the OCR engine used by Gale. This agreement supports being able to use the OCR engine confidence value to accurately assess OCR quality also beyond the small subset.

However, ECCO 1 and ECCO 2 arise from different OCR processes, and the above correlation strictly applies only to ECCO 1 due to the ECCO-TCP only containing material from it. Yet, the confidence scores for both ECCO 1 and ECCO 2 do follow similar patterns with regard to time, language and other secondary axes, suggesting that also the ECCO 2 engine confidence could be trusted.

Importantly though, the confidence estimates of the OCR engine used for digitising ECCO 2 are probably not directly comparable to those reported by the engine used for ECCO 1. To wit, the confidence scores reported by the ECCO 2 process are consistently lower than those reported by the ECCO 1 process. Instead of indicating a general decrease in OCR quality for the publications scanned later, this more likely just means that the confidence estimates operate on different scales overall.

Figure 4 charts the OCR confidence measures in the two subcollections against time. For both collections, median accuracy improves with time, particularly from 1700 to 1750. At the same time, both collections contain many outliers with a remarkably lower confidence. On a surface level, one would also be tempted to draw the conclusion that the quality variation is more intense in ECCO 2, but that may just be an artifact of the different confidence scales used in the two collections.

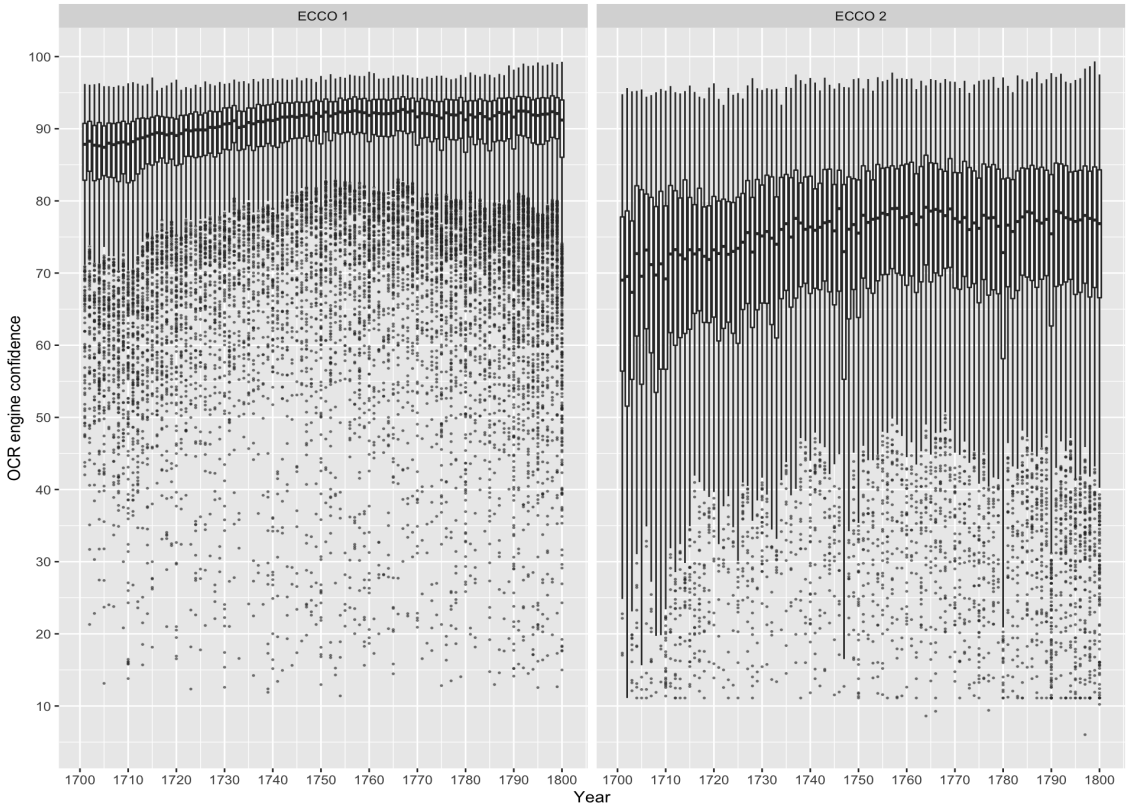

Figure 4: OCR quality in ECCO 1 and ECCO 2 through time[17]

While ECCO is primarily composed of English texts, if one is interested in the small subparts of it which are not, one will be interested in how the OCR quality is affected by the language. First, to verify whether language had an effect on the reliability of the OCR confidence estimates, we calculated the correlation between ECCO-TCP transcriptions and ECCO OCR versions for the different languages. That collection contains only a few documents in languages other than English, including French (N=31) and Welsh (N=94), and 443 documents with an unknown language. The correlation between the manually curated quality (F1 scores) and the automated OCR confidence intervals was 0.8 across all languages without any significant difference. Thus, the confidence scores seem to be trustworthy indicators of OCR quality also for non-English documents.

Expanding from this to look at the OCR confidences across all languages in ECCO (cf. Figure 5), we see that the median confidence is lower for languages other than English. Of particular interest here is that German has a remarkably lower OCR confidence than the other languages. This might be due to the system not being properly configured for German special characters, which do appear in the automated transcriptions, but not as often as they should.

[17] Note that the OCR confidences provided by the engine (vertical axis) are not comparable between ECCO 1 and ECCO 2.

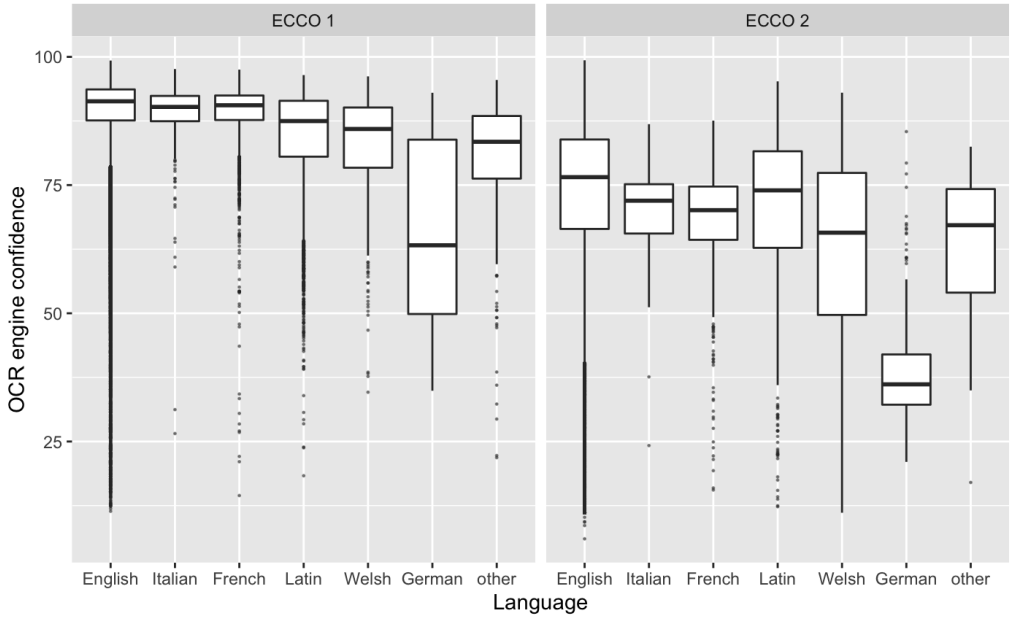

Figure 5: OCR quality in different languages[18]

3. Conclusion

ECCO is a primary source for anyone interested in eighteenth-century English language. The availability of ECCO for text mining is also changing the way scholars work (and will work in the future). While these kinds of big data sources will gain even more prominence in the future, the role of source criticism will be more and more important to all fields that want to use large historical collections. The interests of linguists, historians and data scientists are thus mutual and all these relevant expertises are needed.

Our analysis of ECCO has shown that different kinds of biases in the data are evident based on the general composition of the collection alone. The geographical distribution of ECCO is uneven compared to the full eighteenth-century British printing record. There are also historical reasons for the geographical imbalance but the main reason for the missing documents in English is the process of putting ECCO together. The temporal distribution of ECCO is likewise uneven, with the end of the eighteenth century dominating the corpus mainly because the printing activity was increasing during that time. The reprint activity, too, is higher towards the end of the century, and there are more reprints included during the later eighteenth-century decades in ECCO than earlier. It is also evident that the most popular authors are overrepresented, which

[18] Note that the OCR quality scores (vertical axis) are not comparable between ECCO 1 and ECCO 2.

creates a bias of its own in the reprints that can be detected in the data. The OCR quality in ECCO data is generally remarkably lower compared to results that are achieved in digitisation of scanned sources in the 2020s. Furthermore, the OCR quality is significantly uneven between different parts of ECCO. When we combine this information about the OCR quality with the question of reprints and other issues discussed in our analysis we understand that these biases accumulate. This is visible for example in basic key-word searches. Popular authors are overrepresented already for historical reasons and their presence in ECCO is further amplified due to other biases in the selection process of included works and poor OCR quality. Obviously, the more works you have included in the data, the greater the likelihood of them turning up on different occasions.

There are good reasons why we should take the opportunity to use ECCO when studying language change seriously. ECCO is a remarkable source in spite of the gaps in the data that we have detected. When we combine an understanding of possible bias in the data with the potential of ECCO for data mining, we may formulate more robust approaches to it in our research. There is great potential in ECCO to study language variation and change when we take into consideration the distinction between ‘a corpus as the input for a reader’ (canonical works or ideas) and ‘a corpus as the output of a writer’ (neologisms), the increase in precise definitions of philosophical concepts, and the correlation between OCR engine confidence and quality. For example, the regional variation of eighteenth-century printed English is an aspect that we can study based on this source. But what is needed is the understanding of the historicity of the source both as actual historical processes and also as the layering of a collection that has a complicated provenance. After grasping this historicity, we are then able to think of different ways to limit the effect of these biases.

We believe that investigating language by the use of ECCO is possible, given that careful work is put into taking different aspects into consideration and the research questions are matched with what is possible to do with such a biased and largely inaccurate corpus. Our aim in this article has been to bring forward some crucial limitations of ECCO in the use of corpus linguistics. The next step will be to overcome these limitations, especially with respect to the low OCR quality that renders many

intuitively useful interfaces for modelling ECCO, such as Gale's own *Digital Scholar Lab*, currently virtually unusable for many research tasks.[19]

References

Alston, Robin. 1981. ESTC texts on microfilm. *Factotum: Newsletter of the XVIIIth century STC* 12: 2–3.

Alston, Robin. 2004. The history of ESTC. *The Age of Johnson* 15: 269–329.

Bonnell, Thomas F. 2009. Reprint trade. In Michael F. Suarez and Michael L. Turner eds. *The Cambridge History of the Book in Britain. Vol. V. 1695–1830*. Cambridge: Cambridge University Press, 699–709.

Bullard, Paddy. 2013. Digital humanities and electronic resources in the long eighteenth century. *Literature Compass* 10/10: 748–760.

Cayley, Seth. 2017. Digitization for the masses: Taking users beyond simple searching in Nineteenth-Century Collections Online. *Journal of Victorian Culture* 22/2: 248–255.

Davies, Mark. 2012. Some methodological issues related to corpus-based investigations of recent syntactic changes in English. In Terttu Nevalainen and Elizabeth Closs Traugott eds. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 157–174.

Eighteenth Century Collections Online. https://www.gale.com/intl/primary-sources/eighteenth-century-collections-online

English Short Title Catalogue. http://estc.bl.uk

Gale. 2016. *Eighteenth Century Collections Online*: The most comprehensive online library of English and foreign-language titles printed in the United Kingdom during the eighteenth century, plus thousands of important works printed in English elsewhere. https://www.gale.com/binaries/content/assets/gale-us-en/primary-sources/eighteenth-century-collections-online/ecco-roll-fold-2016-web.pdf

Greenfield, Sayre. 2010. ECCO OCR troubleshooting. *Early Modern Online Bibliography*. https://earlymodernonlinebib.wordpress.com/ecco-ocr-troubleshooting-by-sayre-greenfield/ (15 January, 2020.)

Gregg, Stephen. 2020. *Old Books and Digital Publishing: Eighteenth-Century Collections Online*. Cambridge: Cambridge University Press.

Gries, Stefan Th. 2012. Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: Towards more and more fruitful exchanges. In Joybrato Mukherjee and Magnus Huber eds. *Corpus Linguistics and Variation in English. Theory and Description*. Amsterdam: Rodopi, 41–63.

Hill, Mark J. and Simon Hengchen. 2019. Quantifying the impact of dirty OCR on historical text analysis: *Eighteenth Century Collections Online* as a case study. *Digital Scholarship in the Humanities* 34/4: 825–843.

Hiltunen, Turo, Joe McVeigh and Tanja Säily. 2017. How to turn linguistic data into evidence? In Turo Hiltunen, Joe 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 April, 2021.)

[19] https://www.gale.com/intl/primary-sources/digital-scholar-lab

Hitchcock, Tim and Robert Shoemaker. 2007. The value of the proceedings as a historical source. *Old Bailey Proceedings Online*. https://www.oldbaileyonline.org/static/Value.jsp (16 April, 2021.)

Ijaz, Ali, Leo Lahti, Iiro Tiihonen and Mikko Tolonen. 2019. Analytical determination of editions from bibliographic metadata. In Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi eds. *Proceedings of the Research Data and Humanities 2019 Conference: Data, Methods and Tools*. Oulu: University of Oulu. http://urn.fi/urn:isbn:9789526223216 (24 April, 2021.)

Kinley, Welly. 2003. Digital ECCOs of the eighteenth century. *eContent*, November Issue. https://chnm.gmu.edu/digitalhistory/links/pdf/introduction/0.27b.pdf (24 April, 2021.)

Lahti, Leo, Niko Ilomäki and Mikko Tolonen. 2015. A quantitative study of history in the *English Short-Title Catalogue* (ESTC) 1470–1800. *LIBER Quarterly* 25/2: 87–116.

Lahti Leo, Eetu Mäkelä and Mikko Tolonen. 2020. Quantifying bias and uncertainty in historical data collections with probabilistic programming. In Folgert Karsdorp, Barbara McGillivray, Adina Nerghes and Melvin Wevers eds. *Proceedings of the Workshop on Computational Humanities Research 2020.* Aachen: CEUR-WS.org, 280–289.

Lahti, Leo, Jani Marjanen, Hege Roivainen and Mikko Tolonen. 2019. Bibliographic data science and the history of the book (c. 1500–1800). *Cataloging & Classification Quarterly* 57/1: 5–23.

Linguistic DNA. 2017. Experimenting with the imperfect: ECCO & OCR. https://www.linguisticdna.org/ecco-ocr/ (20 February, 2020.)

Mäkelä, Eetu, Krista Lagus, Leo Lahti, Tanja Säily, Mikko Tolonen, Mika Hämäläinen, Samuli Kaislaniemi and Terttu Nevalainen. 2020. Wrangling with non-standard data. In Sanita Reinsone, Inguna Skadiņa, Anda Baklāne and Jānis Daugavietis eds. *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference 2020*. Aachen: CEUR-WS.org, 81–96.

Tolonen, Mikko, Eetu Mäkelä and Leo Lahti. 2017. Analysing eighteenth-century key-terms and phrases using ECCO and ESTC. *Paper presented at the British Society for Eighteenth Century Studies BSECS 46th Annual Conference*, Oxford.

Tolonen, Mikko, Leo Lahti, Jani Marjanen and Hege Roivainen. 2018. A quantitative approach to book-printing in Sweden and Finland, 1640–1828. *Historical Methods* 52/1: 57–78.

Tolonen Mikko, Mark Hill, Ali Ijaz, Ville Vaara and Leo Lahti. 2021. Examining the early modern canon: The *English Short Title Catalogue* and large-scale patterns of cultural production. In Ileana Baird ed. *Data Visualization in Enlightenment Literature and Culture*. London: Palgrave Macmillan, 63–119.

Corresponding author
Mikko Tolonen
University of Helsinki
P.O. Box 24
00014. Helsinki
Finland
Email: mikko.tolonen@helsinki.fi

received: March 2020
accepted: April 2021

RiCL Research in Corpus Linguistics

Challenges of releasing audio material for spoken data: The case of the *London-Lund Corpus 2*

Nele Põldvere[a/b] – Johan Frid[a] – Victoria Johansson[a] – Carita Paradis[a]
Lund University[a] / Sweden
University of Oslo[b] / Norway

Abstract – This article aims to describe key challenges of preparing and releasing audio material for spoken data and to propose solutions to these challenges. We draw on our experience of compiling the new *London-Lund Corpus 2* (LLC-2), where transcripts are released together with the audio files. However, making the audio material publicly available required careful consideration of how to, most effectively, 1) align the transcripts with the audio and 2) anonymise personal information in the recordings. First, audio-to-text alignment was solved through the insertion of timestamps in front of speaker turns in the transcription stage, which, as we show in the article, may later be used as a valuable complement to more robust automatic segmentation. Second, anonymisation was done by means of a *Praat* script, which replaced all personal information with a sound that made the lexical information incomprehensible but retained the prosodic characteristics. The public release of the LLC-2 audio material is a valuable feature of the corpus that allows users to extend the corpus data relative to their own research interests and, thus, broaden the scope of corpus linguistics. To illustrate this, we present three studies that have successfully used the LLC-2 audio material.

Keywords – audio-to-text alignment; anonymisation; corpus compilation; spoken corpora; prosody; *Praat*

1. INTRODUCTION[1]

With the advent of several new spoken corpora, challenges related to the various aspects of spoken corpus compilation are currently receiving more and more attention in the research community (e.g., Andersen 2016; Diemer *et al*. 2016; Kirk 2016; Sauer and

[1] We would like to express our gratitude to Bas Aarts and Sean Wallis from the *Survey of English Usage* (University College London) for giving us access to the LLC-1 audio material, and to the two anonymous reviewers, the editors of this special issue, and the general editors of *RICL* for their insightful comments on an earlier version of the manuscript. We are also grateful to *Lund University Humanities Lab*. This work has in part been funded by an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2019–2024; contract no. 2017-00626). The compilation of LLC-2 has largely been funded by the *Linnaeus Centre for Thinking in Time: Cognition, Communication, and Learning*, financed by the Swedish Research Council (grant no. 349-2007-8695), and the Erik Philip-Sörensen Foundation.

Lüdeling 2016; Weisser 2017). However, these studies tend to focus on the part of corpora that constitutes the machine-readable data for spoken corpus research, that is, the transcriptions, rather than on the primary data from which the transcriptions have been derived, that is, the original audio recordings. The aim of this article is to describe and propose solutions to key challenges of preparing and releasing audio material for spoken data. It is based on our experience of compiling the new *London-Lund Corpus 2* (LLC-2; Põldvere *et al*. in press b.; see also the user guide in Põldvere *et al*. in press a.). LLC-2 is a half-a-million-word corpus of spoken British English dating from 2014 to 2019, and its compilation followed the same design criteria as in the world's first spoken corpus, the *London-Lund Corpus of Spoken English* (LLC-1) with data from the 1950s to the 1980s (see Section 3.1). In contrast to many other widely used spoken corpora in English, the transcripts in LLC-2 are released together with the audio files. However, for this to be possible, we had to tackle two major challenges: 1) the alignment of the transcripts with the audio files and 2) the anonymisation of personal information in the recordings. First, audio-to-text alignment was necessary in order to allow users to easily find relevant sections of the transcripts in the audio files and to improve the usability of LLC-2. The choice was between sophisticated automatic segmentation techniques and the simpler alternative of inserting timestamps during transcription. In this article, we explain why we decided to opt for the latter option and demonstrate the feasibility of combining it with more robust automatic segmentation (see Section 3.2). Second, the anonymisation of the audio recordings was mandatory out of respect for the speakers' privacy and legal protection of personal data. This procedure was, however, not straightforward because it required careful manipulation of the speech signal. We describe and explain why and how we anonymised the LLC-2 audio recordings using a *Praat* script developed by Hirst (2013) (see Section 3.3).

The benefits of releasing the LLC-2 audio material to the research community are immense. As is the case in many other spoken corpora, the transcriptions in LLC-2 are orthographic and contain information about basic features of spoken interaction such as pauses, overlapping speech and nonverbal vocalisations, but not prosodic and temporal information about pitch movement and the length of transitions between speaker turns. These features are, however, important for spoken language research because they carry useful information about speaker intent. Moreover, having access to prosodic and temporal information about speech broadens the field of corpus linguistics to go beyond

the traditional areas of lexicology, morphology, syntax and discourse analysis. With the release of the LLC-2 audio material, users can pursue these interests and extend the transcriptions using different speech analysis and annotation tools. To illustrate this, we provide examples of previous research on data from LLC-2 where the audio material was successfully used to carry out prosodic and temporal investigations of spoken interaction (see Section 3.4). Section 2 provides the background information.

2. AUDIO MATERIAL IN SPOKEN CORPORA

In this section, we will first present the core practices of how speech is represented in spoken corpora, and how these practices have influenced research conducted in two areas of linguistic inquiry: prosody and turn-taking (Section 2.1). Then, we review five well-known corpora of spoken British English and the extent to which they have made available the original audio material to facilitate more thorough investigations of the prosodic and temporal aspects of spoken interaction (Section 2.2).

2.1. Representations of speech in corpus linguistics

Compiling a spoken corpus is a complex and time-consuming task that requires careful decision-making at each stage of the process. Perhaps the most well-documented stage is the transcription stage, where the speech is turned into written form to provide the machine-readable material for browsing, searching and counting chunks in the corpus (e.g., Ochs 1979; Du Bois 1991; Crowdy 1994; Edwards 1995; Andersen 2016). To add value, the transcriptions may be complemented with layers of markup and annotation that convey additional information about the original speech event (e.g., Edwards 1995; Leech 2004; Kirk 2016; Sauer and Lüdeling 2016; Gries and Berez 2017; Weisser 2017). While corpus markup contains information about structural features inherent in speech production —such as who speaks, when and for how long— the function of corpus annotation is to add to the transcriptions linguistic information about, for example, parts-of-speech and syntactic parsing (Kirk and Andersen 2016: 291–292). The level of detail of the transcription, markup and annotation schemes adopted in spoken corpus projects depends on, among many other factors, the intended future uses of the corpus. Most of these uses tend to fall into the traditional areas of corpus linguistics such as lexicology, morphology, syntax and discourse analysis.

A much less well-documented stage of spoken corpus compilation is the process of making available the primary data from which the transcriptions have been derived, namely the original audio recordings (see, however, Diemer *et al*. 2016; Sauer and Lüdeling 2016; Schmidt 2016; Hoffmann and Arndt-Lappe submitted). This stage is, however, important because even the most detailed transcription, markup and annotation schemes lose valuable information about the original speech event in the transfer of the data to written form. Thus, the release of the audio material alongside the transcripts has the potential of extending corpus linguistics in new directions, that is, where the exploration of additional spoken features can add to our understanding of how spoken interaction works. In this article, we focus on two areas where this may prove useful: prosody and turn-taking.

Prosody is an essential component of human communication. Every utterance in spoken interaction contains prosodic features that convey important information about speaker intent. For example, the same expression has different interpretations depending on whether it receives a falling or rising intonation (compare *r\ight* as an expression of agreement and *r/ight* as a confirmation-seeking question).[2] Prosody research draws on data either from controlled laboratory experiments or speech corpora designed specifically for prosodic analyses (e.g., the *IViE Corpus of English Intonation in the British Isles*; see Grabe 2004).[3] Accordingly, the availability and quality of audio files are of utmost importance as "the research for which they are used is frequently focused on the speech signal itself" (Wichmann 2008: 188). This is different from corpus linguistics where, normally, corpora are intended to be useful for a wide variety of linguistic interests, and where many researchers consider the primary data to be the transcriptions with annotations of lexical, morpho-syntactic and discourse features (Oostdijk and Boves 2008: 196).

Turn-taking is a basic mechanism of dialogic spoken interaction and one of the main foci of Conversation Analysis (CA). Similar to corpus linguists, conversation analysts base their analyses on recordings of naturally occurring speech; however, most conversation analysts collect and transcribe their own data (Hoey and Kendrick 2017: 155) in order to ensure that the transcriptions are detailed enough to permit meaningful analyses for their purposes. For example, CA transcripts contain detailed information

[2] In the first instance, \ indicates a falling intonation contour from a high accented syllable and, in the second instance, / indicates a rising intonation contour from a low accented syllable.
[3] http://www.phon.ox.ac.uk/files/apps/IViE

about the boundaries of overlapping speech and the length of gaps between speaker turns in milliseconds. This information is important for understanding speaker intent because turns produced after a noticeable gap (after, say, 600 ms) have been found to signal interactional trouble (Roberts *et al*. 2006) and may be interpreted as "the first move toward some form of disagreement/rejection" (Clayman 2002: 235). The level of detail needed to transcribe the recordings means that the datasets in CA are relatively small, which goes well with the qualitative focus of the framework. More recent quantitative work, however, has also consulted larger corpora. Roberts *et al*. (2015), for example, used the *NXT-format Switchboard Corpus* (Calhoun *et al*. 2010), which includes detailed temporal chunking of phonetic segments and words, to automatically estimate the duration of transitions between speaker turns. Yet other quantitative studies in CA have made use of various speech analysis and annotation tools to manually identify beginnings and ends of speaker turns (e.g., *Praat* in Kendrick and Torreira 2015). Thus, analyses of the organisation of turn-taking in spoken interaction rely heavily on the availability either of richly annotated transcripts or the original audio material or both. However, as we will show in Section 2.2, it is not common that these features are available in spoken corpora, let alone the possibility to combine the transcripts with the audio to facilitate even more thorough analyses of turn-taking and prosody in spoken interaction.

2.2. A review of corpora of spoken British English

In this section, we review five well-known corpora of spoken British English and the extent to which they give access to the original audio material. The corpora are: 1) the spoken component of the first *British National Corpus* (Spoken BNC1994; cf. BNC Consortium 2007),[4] 2) the spoken component of the second *British National Corpus* (Spoken BNC2014; cf. Love *et al.* 2017),[5] 3) the *British Component of the International Corpus of English* (ICE-GB; cf. Nelson *et al.* 2002),[6] 4) the first *London-Lund Corpus* (LLC-1; Greenbaum and Svartvik 1990)[7] and 5) the second *London-Lund Corpus* (LLC-2; cf. Põldvere *et al*. in press b.).[8] Spoken BNC1994 and Spoken BNC2014 are

[4] http://www.natcorp.ox.ac.uk
[5] http://corpora.lancs.ac.uk/bnc2014
[6] http://ice-corpora.net/ice/index.html
[7] http://icame.uib.no
[8] https://projekt.ht.lu.se/llc2

large, multi-million-word corpora recorded in the early 1990s and 2010s, respectively. The remaining corpora are considerably smaller with approximately half-a-million words each. ICE-GB contains data from the 1990s, while LLC-1 was recorded as early as in the 1950s–1980s and LLC-2 was recorded as recently as 2014–2019. The corpora were selected for the review because they all provide access to spontaneous everyday conversation (either as part of the corpus or in full), which is the most rewarding conversational setting for studies of prosody and turn-taking, and they are available either for free or after payment of a licence fee.

Table 1 below presents basic information about how the corpora were transcribed, marked up and annotated to facilitate prosodic and temporal analyses of spoken interaction, and the availability of audio material in the corpora. The idea is to determine whether users can carry out analyses of the topics if they only have access to the transcripts, and, if not, what options there are for them to consult the original audio recordings.

As can be seen in Table 1, the general approach to transcription in the corpora is to adopt an enhanced orthographic transcription scheme, which involves a transcription of words enhanced by markups and annotations of basic spoken features such as pauses, overlapping speech, nonverbal vocalisations (e.g., laughter), etc. However, most of the corpora (i.e. Spoken BNC1994 and, to a lesser extent, Spoken BNC2014) contain only limited prosodic annotation, such as rough indications of pitch contours, or none at all (ICE-GB[9] and LLC-2). The main reasons why orthographic transcriptions take precedence in spoken corpora are because they are easier and less costly to implement than prosodic transcriptions, and because orthographic transcriptions are sufficient for a wide variety of corpus linguistic studies (Atkins *et al*. 1992: 10; Love *et al*. 2017: 334).

[9] It should be noted that *Systems of Pragmatic Annotation in the Spoken Component of ICE-Ireland* (SPICE-Ireland; cf. https://johnmkirk.etinu.net/cgi-bin/generic?instanceID=11), the pragmatically annotated version of the Irish component of the *International Corpus of English*, has been annotated for pitch location and direction (Kirk 2016).

Corpus	Transcription, markup and annotation			Audio material
	General	Prosody	Turn-taking	
Spoken BNC1994 (10 million words).	Enhanced orthographic transcription.	Little prosodic annotation (e.g., question marks are used to indicate *questioning utterances*).	Distinction between short (<5s) and long gaps; boundaries, but not length, of overlaps are marked.	Downloadable WAV files available from Audio BNC for free; audio playback of query matches available from the free online interface BNCweb; not all recordings included; subset of the recordings published on *Corpuscle*[10] (cf. Meurer 2012) as part of *The Bergen Corpus of London Teenage Language* (COLT), cf. Stenström *et al.* (1998).[11]
Spoken BNC2014 (11 million words).	Enhanced orthographic transcription.	Only questions with obvious rising intonation are marked.	Distinction between short (<5s) and long gaps; only presence/absence of overlaps is marked.	No public access to audio material; plans to anonymise and release the recordings.
ICE-GB (600,000 words).	Enhanced orthographic transcription.	No prosodic annotation.	Distinction between short (one syllable) and long gaps; boundaries, but not length, of overlaps are marked.	Audio playback of the recordings available at a cost from the *UCL Survey of English Usage*.
LLC-1 (500,000 words).	Prosodic and paralinguistic transcription.	Extensive prosodic annotation (e.g., tone units, nuclear tones, stress).	Distinction between short (one syllable) and long gaps; boundaries, but not length, of overlaps are marked.	No public access to audio material.
LLC-2 (500,000 words).	Enhanced orthographic transcription.	No prosodic annotation.	Only one type of gap is included (one syllable or longer); boundaries, but not length, of overlaps are marked.	Downloadable WAV files available from the *Lund University Humanities Lab*'s corpus server; all recordings included.[12]

Table 1: The comparison of the nature of transcriptions and the availability of audio material of five well-known corpora of spoken British English

The only corpus in Table 1 that contains detailed prosodic and paralinguistic transcriptions is LLC-1. The corpus is annotated for prosodic features such as tone unit boundaries, the direction of the nuclear tone, varying degrees of stress, and paralinguistic features such as whisper and creak (Svartvik and Quirk 1980; Greenbaum and Svartvik 1990). The prosodic annotations have provided searchable data for a broad range of corpus linguistic studies (e.g., Stenström 1984; Aijmer 1996; Paradis 1997; Altenberg 1998; Lenk 1998; Kaufmann 2002; Romero-Trillo 2014; Põldvere *et al.* 2016; Kimps 2018; Lin 2018). However, with data from the 1950s to the 1980s, LLC-1

[10] https://clarino.uib.no/korpuskel/page
[11] http://korpus.uib.no/icame/colt/
[12] Only one 10-minute university lecture is unavailable as per a request from the lecturer.

is less suited for contemporary investigations of speech. This is because prosodic alterations and variants have been found to go hand in hand with meaning shifts and change (Paradis 2008; Wichmann *et al*. 2010; Wichmann 2011; Põldvere and Paradis 2019, 2020), and the prosodic patterns found in English some 50 years ago may not be the same as in contemporary speech. Furthermore, the annotations in LLC-1 are based on auditory analysis, which is heavily reliant on subjective impressions (cf. Wichmann 2008: 202). Therefore, users may want to inspect the original speech signal to reinforce or counter auditory impressions and, thus, obtain more reliable results (see Section 3.4).

Investigations of turn-taking in the corpora in Table 1 are facilitated to the extent that all of them are annotated for whether the transition between the speaker turns is a gap or an overlap.[13] Many of the corpora have made available additional information such as distinctions between short and long gaps, and the boundaries of the overlapping speech, but none of them has gone as far as to measure the length of time between the speaker turns, as is commonly the case in CA (see Section 2.1 above). Thus, Table 1 shows that, while all the corpora facilitate rough analyses of the organisation of turn-taking, they are less well-suited for thorough investigations of the timing of turns in conversation.

When we compare the transcription schemes to the availability of the audio material, it becomes clear that the shortcomings of the transcriptions are not always compensated for by access to the original audio recordings or the access is in some way restricted. This explains, at least partly, why prosody and turn-taking —both of which are heavily dependent on the availability of the original speech signal— are conspicuously under-researched in corpus linguistics. The corpora that do not provide any kind of public access to the audio material are Spoken BNC2014 and LLC-1.[14] The main reason for this is that the recordings have not been anonymised and therefore cannot be publicly released (e.g., Love *et al*. 2017: 335; see also Section 3.3). The ICE-GB audio material is available via audio playback from the *Survey of English Usage* at University College London, which means that users can search for an expression in the

[13] Note that, for current purposes, we use the term ‘gap’ to refer to what are more commonly known in corpus annotation as ‘pauses’; however, they are a special kind of pauses in that they only occur between speaker turns.

[14] According to *UCL Survey of English Usage* (2020), the *Diachronic Corpus of Present-Day Spoken English* (cf. https://www.ucl.ac.uk/english-usage/projects/dcpse/), of which LLC-1 is part, only contains the orthographic transcriptions and not the original audio files. In the early days, researchers had to travel to the *Survey of English Usage* in London to be able to listen to the recordings. Many researchers today have access to the digital files; however, no systematic access has been provided to date.

corpus and listen to the passage containing that expression (*UCL Survey of English Usage* 2020; see also Wallis *et al*. 2006). However, this feature of ICE-GB is only available after payment of a licence fee, which together with the transcripts and the software for searching the corpus may amount to as much as £600–800 for an individual, single-copy licence.[15] Access to the Spoken BNC1994 audio material is free of charge. Moreover, users can choose between two formats: 1) the complete WAV audio files are available for download from Audio BNC (Coleman *et al*. 2012), and 2) the BNCweb online interface allows users to play back, as well as download, the audio of the query match and its immediate context (Hoffmann *et al*. 2008; Hoffmann and Arndt-Lappe submitted). The only downside is that neither Audio BNC nor BNCweb provides access to the complete dataset. According to Coleman *et al*. (2012: para. 2), "[t]here is a substantial number of XML transcription files for which we may no longer have the original audiotapes […] we also have quite a few recordings that we haven't yet related to any transcription." Moreover, for copyright reasons, neither of the audio editions of Spoken BNC1994 gives access to the recordings of a subset of BNC1994, namely COLT, which instead are published on the online interface *Corpuscle* via audio playback (for more information on *Corpuscle*, see Section 4). Coleman *et al*. (2012) estimate the size of the missing dataset in Audio BNC (and, by extension, BNCweb) to be around 2.5 million words. As we will show in Section 3.4, this is enough to pose problems for those who wish to use the audio material in their research.

In our work with the design and compilation of LLC-2, we decided to address the above-mentioned shortcomings and provide access to the complete set of recordings, which are time-aligned with the transcripts and anonymised to adhere to ethical standards (see Section 3 for details). The recordings can be accessed from the *Lund University Humanities Lab*'s corpus server as downloadable WAV files. We decided to make the LLC-2 audio material publicly available to allow users to extend the orthographic transcriptions relative to their own research interests using any of the free software available for annotating and analysing spoken data. However, preparing the audio files for release did not come without its challenges, which are the same challenges that have discouraged or prevented many corpus developers before us from doing it. The next section focuses on how we tackled these challenges and, thus,

[15] The prices are as of April 2021.

facilitated the investigation of prosodic and temporal aspects of spoken interaction in LLC-2 in subsequent research.

3. Challenges of preparing LLC-2 audio files for release

This section presents key challenges of making the LLC-2 audio material available to the research community. After a brief description of LLC-2 in Section 3.1, we examine the steps that we took to overcome two challenges of preparing the LLC-2 audio material for public release, audio-to-text alignment (Section 3.2) and anonymisation (Section 3.3). Section 3.4 presents three studies based on data from LLC-2 that demonstrate the usefulness of making the audio material publicly available.

3.1. LLC-2

As already mentioned in Section 1, LLC-2 is a half-a-million-word corpus of spoken British English dating from 2014 to 2019 (Põldvere *et al.* in press b.; see also the user guide in Põldvere *et al.* in press a.). It covers a range of discourse contexts including private contexts such as face-to-face conversation and phone/CMC conversation,[16] as well as public contexts such as broadcast media, parliamentary proceedings, spontaneous commentary, legal proceedings and prepared speech. In addition, efforts have been made to control for certain demographic categories such as the age and gender of the speakers. The size and design of LLC-2 are comparable to those of LLC-1 with data from the 1950s to the 1980s. As a result, LLC-2 can be used to study naturally occurring contemporary speech, on the one hand, and, on the other hand, it gives researchers the opportunity to make principled diachronic comparisons with LLC-1 of speech over the past half a century (see Section 3.4). The corpus will be released to the research community for free via the *Lund University Humanities Lab*'s corpus server in autumn 2021 (see also Section 4).[17] The release contains, among many other things, 184 XML-formatted transcription files and 183 audio files in WAV format.[18] In order to

[16] CMC = *Computer-Mediated Communication*.
[17] The corpus server can be accessed at https://www.humlab.lu.se/facilities/corpus-server
[18] In general, LLC-2 contains 100 texts, each around 5,000 words in size, with corresponding audio recordings, but since one text in the corpus can contain material from one recording only, or it can consist of multiple shorter recordings revolving around a similar subject matter and/or involving the same speaker(s), the total number of transcription and audio files is considerably higher.

facilitate the release of the audio material, we had to tackle two key challenges, which are discussed in the next two sections.

3.2. Audio-to-text alignment

The first key challenge was the alignment of the transcripts with the recordings. Audio-to-text alignment of this kind involves linking particular sections in the transcripts to the corresponding locations in the recordings in order to enhance the usability of the corpus. There are two broad options for how to deal with this (Thompson 2004). On the one hand, corpus developers may use highly sophisticated procedures for automatic alignment, which yield a best-fitting phonetic transcription of the audio and provide detailed timing information about all the vowels, consonants and words in the recordings. Such an approach was adopted in Spoken BNC1994, both in Audio BNC and BNCweb (Coleman *et al*. 2012; Hoffmann and Arndt-Lappe submitted). On the other hand, a simpler solution is to manually place markers in the transcripts to point to precise timings in the audio files. This functionality is often built into transcription software (e.g., ELAN; see Wittenburg *et al.* 2006) and it gets integrated into the transcription stage. In LLC-2, we adopted the latter approach. The reason for this was that the insertion of timestamps is easy to implement and provides sufficiently accurate points of entry into the audio files for a wide variety of corpus linguistic studies.

The tool used to insert timestamps in LLC-2 was *InqScribe* (2005–2020). *InqScribe* is a low-cost transcription software tool that enables users to perform all their transcriptions and audio playback in the same window. An important feature of the software is that it includes a simple functionality for inserting timestamps by means of customised keyboard shortcuts. In LLC-2, the insertion of timestamps was administered on a turn-by-turn basis. This means that, at the onset of each speaker turn in the recordings, a customised keyboard shortcut was used to launch a snippet containing the timestamp and the speaker's unique identifier. In recordings with only one speaker (e.g., prepared speech) or recordings with overly long contributions by one speaker (e.g., spontaneous commentary), timestamps were inserted every minute. The combination of the timestamps with the speakers' unique identifiers, inserted with one keyboard shortcut, meant that no extra time had to be spent on inserting the timestamps separately. Thus, this technique can be scaled up to larger corpora containing spontaneous everyday conversation, which, due to its messiness, still requires manual transcription (see McEnery 2018).

In order to facilitate compatibility with existing corpus tools, the *InqScribe* files were converted into canonical XML files. XML works on the principle that whatever is enclosed within angle brackets is treated as corpus markup and whatever falls outside the angle brackets is the actual corpus text. Following the recommendations in Hardie (2014), we made additions to the standard set of XML tags where required. This is illustrated in the XML transcript in Figure 1 below, where each speaker turn is enclosed within the <turn> tag, which attributes for the number of the turn (*n*), the timestamp with the value format hh:mm:ss.ms, and, finally, the unique speaker identifier (*who*). The timestamps in LLC-2 help users find the appropriate places in the recordings with minimal effort, thus serving as valuable points-of-entry for more thorough analyses of the speaker turns. An obvious shortcoming of the XML transcripts is that they do not allow for immediate audio playback of the turns; however, we will facilitate this through the release of LLC-2 from an online interface (see Section 4 for details).

The availability of both the orthographic transcriptions and the corresponding audio recordings in LLC-2 also allows for the implementation of more sophisticated automatic alignment techniques to extend the use of the corpus to more areas. For example, for phonetic research it is usually desirable to have phonetic transcriptions as well as phonetically time-aligned boundaries between segments (Yuan *et al*. 2018). With a project of this scale, manual segmentation is not feasible as it is very costly in people-hours. Instead, automatic segmentation may be obtained through forced alignment. Forced alignment is the process of automatic alignment of an audio recording to a given transcript. Currently, the best systems for forced alignment make use of language-dependent dictionaries and acoustic models (Hosom 2009). The dictionaries are used to look up canonical phonetic representations of the words in the transcript, and the pre-trained acoustic models contain statistical representations of the acoustic information of the phonemes in language. The acoustic models analyse the audio recording, and the result is matched with the phonetic representation obtained from the dictionary in order to produce time-aligned segmentation. Some researchers have reported that there is a small decrease in accuracy compared to manual alignment (e.g., Hosom 2000). However, it is also the case that manual alignment by humans introduces a degree of random variability, while automatic alignment is rigorously systematic (see, e.g., Cosi *et al*. 1991; Baghai-Ravay *et al*. 2009). Weighing this in, the time gained from using automatic alignment is worth it.

```
<turn n="1" timestamp="00:00:00.17" who="S004">it's like some fitness place</turn>
<turn n="2" timestamp="00:00:02.15" who="S005">oh</turn>
<turn n="3" timestamp="00:00:02.25" who="S004">and some woman was just handing them out <pause/> but it looks alright they <overlap pos="start" n="1"/>do like <trunc>b</trunc><overlap pos="end" n="1"/></turn>
<turn n="4" timestamp="00:00:06.00" who="S005"><overlap pos="start" n="1"/><trunc>i</trunc> is it<overlap pos="end" n="1"/> all weird juices</turn>
<turn n="5" timestamp="00:00:07.24" who="S004">no they do like banana smoothies and stuff</turn>
<turn n="6" timestamp="00:00:10.19" who="S005">so yeah just weird juices <pause/> <overlap pos="start" n="2"/><vocal desc="laughs"/><overlap pos="end" n="2"/> well this could be nice</turn>
<turn n="7" timestamp="00:00:13.05" who="S004"><overlap pos="start" n="2"/><vocal desc="laughs"/><overlap pos="end" n="2"/></turn>
<turn n="8" timestamp="00:00:15.02" who="S004">any free stuff is <pause/> like I was just passing the woman with the fliers <pause/> and everyone was passing her by <pause/> and then I saw the flier said free and I was
<turn n="9" timestamp="00:00:23.10" who="S005">give <pause/></turn>
<turn n="10" timestamp="00:00:24.11" who="S004">yes <vocal desc="laughs"/> I'm <overlap pos="start" n="3"/>interested now<overlap pos="end" n="3"/></turn>
<turn n="11" timestamp="00:00:26.14" who="S005"><overlap pos="start" n="3"/><vocal desc="laughs"/><overlap pos="end" n="3"/> <pause/> Russell Square is it <overlap pos="start" n="4"/><vocal desc="gasps"/><overlap pos="e
<turn n="12" timestamp="00:00:30.22" who="S004"><overlap pos="start" n="4"/>yeah<overlap pos="end" n="4"/></turn>
<turn n="13" timestamp="00:00:34.02" who="S004"><overlap pos="start" n="5"/>no <vocal desc="laughs"/><overlap pos="end" n="5"/> <pause/> so now I have two free drinks I have the one from Eat which I haven't used yet <pa
<turn n="14" timestamp="00:00:44.00" who="S005">and the Waitrose one <pause/></turn>
<turn n="15" timestamp="00:00:45.03" who="S004">I haven't <overlap pos="start" n="6"/>got my Waitrose card yet<overlap pos="end" n="6"/></turn>
<turn n="16" timestamp="00:00:45.13" who="S005"><overlap pos="start" n="6"/>oh if if if you buy one<overlap pos="end" n="6"/> <pause/></turn>
<turn n="17" timestamp="00:00:47.19" who="S004"><vocal desc="laughs"/></turn>
<turn n="18" timestamp="00:00:50.12" who="S005">I'll catch up <pause/></turn>
<turn n="19" timestamp="00:00:51.20" who="S004">and <pause/> this one <pause/></turn>
<turn n="20" timestamp="00:00:55.20" who="S005">what do you have on tomorrow <pause/> you said you were really busy <pause/></turn>
<turn n="21" timestamp="00:00:58.15" who="S004">yes <pause/> I have <pause/> I have Icelandic <pause/> and then Spanish <pause/> so I have grammar for four hours <pause/></turn>
<turn n="22" timestamp="00:01:11.00" who="S005"><overlap pos="start" n="7"/>ugh<overlap pos="end" n="7"/></turn>
<turn n="23" timestamp="00:01:12.02" who="S004"><overlap pos="start" n="7"/>from nine till one<overlap pos="end" n="7"/></turn>
<turn n="24" timestamp="00:01:15.18" who="S005">you're gonna have to have a big lunch <pause/> are you gonna have time for breakfast with her <pause/></turn>
<turn n="25" timestamp="00:01:19.19" who="S004">I don't know <pause/></turn>
<turn n="26" timestamp="00:01:20.25" who="S005">cause you need to eat before that</turn>
<turn n="27" timestamp="00:01:22.03" who="S004">I missed it last time because I <overlap pos="start" n="8"/>was<overlap pos="end" n="8"/> late</turn>
<turn n="28" timestamp="00:01:23.24" who="S005"><overlap pos="start" n="8"/>what<overlap pos="end" n="8"/></turn>
<turn n="29" timestamp="00:01:24.14" who="S005">you missed it this morning</turn>
<turn n="30" timestamp="00:01:25.21" who="S004">no <overlap pos="start" n="9"/>I missed<overlap pos="end" n="9"/> it last <overlap pos="start" n="10"/>Thursday<overlap pos="end" n="10"/></turn>
<turn n="31" timestamp="00:01:26.10" who="S005"><overlap pos="start" n="9"/>oh<overlap pos="end" n="9"/></turn>
<turn n="32" timestamp="00:01:27.03" who="S005"><overlap pos="start" n="10"/>oh last last<overlap pos="end" n="10"/> time you didn't okay <pause/></turn>
<turn n="33" timestamp="00:01:31.18" who="S004">but then I ended up meeting her anyway</turn>
<turn n="34" timestamp="00:01:33.13" who="S005">uhuh <pause/> you don't have any like <trunc>biscuit</trunc> do you have any biscuits left</turn>
<turn n="35" timestamp="00:01:39.00" who="S004">yeah I have three shortbread <vocal desc="laughs"/> biscuits <overlap pos="start" n="11"/>left <pause/> I ate<overlap pos="end" n="11"/> all the rest of them</turn>
<turn n="36" timestamp="00:01:40.26" who="S005"><overlap pos="start" n="11"/><vocal desc="laughs"/> well<overlap pos="end" n="11"/></turn>
<turn n="37" timestamp="00:01:43.10" who="S005">aw well there's your breakfast <pause/> have <pause/> shortbread with Nutella and peanut butter and stuff like that</turn>
<turn n="38" timestamp="00:01:48.28" who="S004">for lunch I made <pause/> boiled eggs <overlap pos="start" n="12"/>eventually<overlap pos="end" n="12"/></turn>
<turn n="39" timestamp="00:01:52.00" who="S005"><overlap pos="start" n="12"/>mhm<overlap pos="end" n="12"/> <pause/> how many <pause/></turn>
<turn n="40" timestamp="00:01:54.13" who="S004">two <pause/></turn>
<turn n="41" timestamp="00:01:55.08" who="S005">mm <pause/></turn>
<turn n="42" timestamp="00:01:56.03" who="S004">I had to go down to the <pause/> to the kitchen on the next floor down <pause/></turn>
<turn n="43" timestamp="00:02:00.27" who="S005">how was it <pause/> was it like amazing <pause/></turn>
<turn n="44" timestamp="00:02:03.15" who="S004">well <pause/> the stuff worked <overlap pos="start" n="13"/>so that<overlap pos="end" n="13"/> was a change <vocal desc="laughs"/></turn>
<turn n="45" timestamp="00:02:05.25" who="S005"><overlap pos="start" n="13"/>mm<overlap pos="end" n="13"/></turn>
```

Figure 1: The illustration of an XML-formatted file in LLC-2

To illustrate the feasibility of forced alignment in LLC-2, we used the WebMAUS system (Schiel 1999; Kisler *et al*. 2017) to produce an alignment of the first few lines of the transcript in Figure 1 above and its corresponding audio recording (see also Sauer and Lüdeling 2016). The transcript and the recording are of a private and spontaneous face-to-face conversation. The result of the WebMAUS system is a TextGrid file, which can be used in the phonetics software *Praat* (Boersma 2001). This is illustrated in Figure 2 where the segmentations have been performed both at the level of words (upper annotation tier) and sounds (lower annotation tiers).

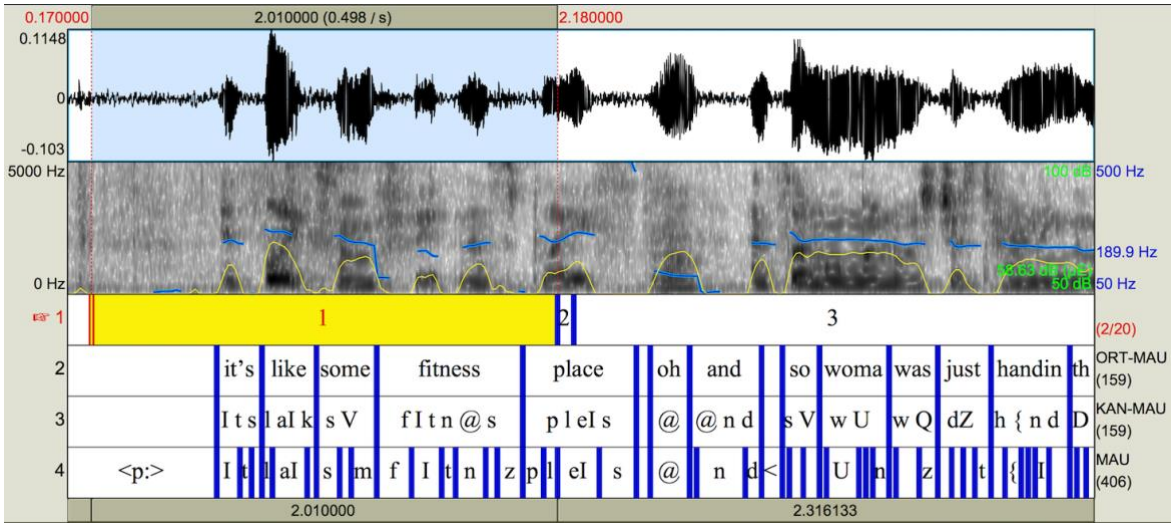

Figure 2: The output of the WebMAUS segmentation system in *Praat* based on the first few lines of the transcript in Figure 1

Looking at it qualitatively, the alignment of the speech signal and the phonetic segments in Figure 2 is very good. Admittedly, there are some misalignments for severely reduced and hasty speech, but that is to be expected in data of this kind. No quantitative evaluation has been made at this stage, as we have no ground truth data to evaluate the alignment against. One could use the automatic alignment as input for a manual correction procedure, which would be much faster than doing full transcription from scratch. Furthermore, the original timestamps in LLC-2 could be used to guide and improve manual editing of the segments. This may prove particularly useful in dealing with overlapping speech and background noise, which are notoriously difficult cases for forced alignment systems. Forced alignment is also highly sensitive to poor audio quality. LLC-2, too, contains private recordings that have been captured with speakers' personal smartphones (e.g., face-to-face conversation) or computer software (e.g., video conversation), which provide audio quality that is far from what phoneticians would consider ideal conditions for forced alignment. This said, we estimate that most of the

data in LLC-2 have been recorded with high-quality digital voice recorders, a feature that we expect to lead to a sufficiently high degree of segmentation accuracy. The alignment in Figure 2 (a private and spontaneous everyday conversation) is a case in point. Thus, looking forward, the prospect of generating for phonetic research automatic transcriptions in LLC-2 seems very promising.

3.3. Anonymisation

The second key challenge that we had to overcome when preparing the LLC-2 audio material for public release was the anonymisation of personal information in the recordings. Anonymisation is mandatory for any publicly available spoken corpus out of respect for the speakers' privacy in line with the *European Union's General Data Protection Regulation* (GDPR). It concerns the removal of all personal information that would allow an individual to be identified. In LLC-2, each speaker was assigned a unique identifier (e.g., <who="S004"> in Figure 1 above) and any references to people's names, addresses, phone numbers, etc., were removed, irrespective of whether these concerned the speakers themselves or any third parties not present in the conversation. The anonymisation was carried out on recordings obtained from private contexts, including 47 texts of face-to-face conversations, nine texts of phone/CMC conversations and two texts of university lectures, but no anonymisation was carried out on radio phone-ins or other types of recordings obtained from the public domain (e.g., podcast discussions).

The anonymisation of personal information during the transcription stage is relatively straightforward. In LLC-2, the transcribers were instructed to mark up all pieces of personal information by enclosing them within the <anon> tag, and to change the information while retaining the word class and number of syllables of the original (e.g., <anon>John</anon> for Sam). In this way, we were able to at least partly retain the socio-cultural information conveyed by the original proper name, including gender and, at times, also ethnicity (see Hasund 1998). A similar procedure was followed in the anonymisation of the transcriptions in ICE-GB and LLC-1 (e.g., Nelson 2002: 7).

The anonymisation of personal information in the original audio recordings is considerably more challenging. It requires careful manipulation of the speech signal, which, in turn, requires special training and adds considerably to the time and money

needed to release the corpus. For example, the reason why the Spoken BNC2014 audio material has not been publicly released yet is because the cost of anonymising the audio recordings went beyond the funding available for the project. However, additional funding will be sought to facilitate this in the future (Love *et al*. 2017: 335). Furthermore, the anonymisation techniques adopted in other spoken corpora have not been completely satisfactory, because they either make certain types of analyses impossible or they pose ethical problems. For example, the approach taken in Spoken BNC1994 consisted of locating and muting the portions of the audio recordings corresponding to the anonymisation tag. Such an approach, however, removes important prosodic information about the original speech signal. Other techniques retain the prosodic information but are problematic in ethical terms. Hirst (2013), for example, reviews two techniques commonly used in psycho-acoustic experiments: 1) the inversion of the spectrum of the speech signal, and 2) the application of a filter that removes the spectral information. However, the problem with those solutions is that, in the first instance, the second inversion of the spectrum restores the original speech signal, and, in the second instance, even quite severe filtering does not make the speech signal unintelligible.

The technique adopted in LLC-2 is based on a *Praat* script written and developed by Hirst (2013).[19] To the best of our knowledge, it has not been implemented in other similar corpora so far.[20] The script works on the basis that the portion of the speech signal that has been marked by the corpus developer with the keyword *buzz* is replaced by a *hum* sound that makes the lexical content of the signal incomprehensible but retains the pitch and intensity envelope of the original. The advantage of this technique is that it is reliable and retains linguistically useful information such as prosody. Moreover, running the script is relatively easy and can be achieved with only minimal training in *Praat*. A somewhat fortuitous side effect is that the task effectively produces data for building a named entity recognition system that can automatically find new portions (names, locations, etc.) that are possible candidates for being anonymised.

An illustration of how the *Praat* script works is given in Figures 3 and 4. Both figures represent the speech signal of a public recording in LLC-2, together with the

[19] The script is freely available at https://hdl.handle.net/11403/sldr000526/v6

[20] The script is currently used in *LangAge Corpora* (cf. http://www.uni-potsdam.de/langage/); however, the corpora are in French and contain specialised content of sociolinguistic interviews with elderly speakers only (Gerstenberg *et al*. 2017).

location and direction of the pitch contour (blue line) and the intensity profile (yellow line).[21] The audio snippet extracted from the recording contains the utterance *Jenni Rodd is a cognitive psychologist at University College London* in which the personal pieces of information are the name and workplace of the person talked about. In Figure 3, this information is marked with the keyword *buzz* to indicate the portions of the speech signal that will be anonymised. Figure 4 presents the end result where the information has been anonymised, and where the pitch and intensity envelopes are the same as in the original.

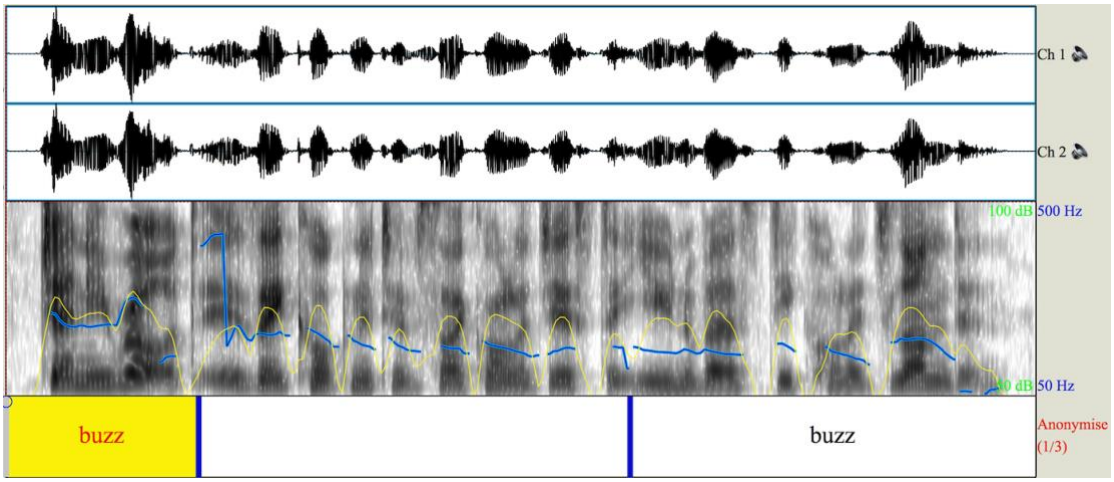

Figure 3: The original speech signal, pitch contour and intensity profile of the utterance *Jenni Rodd is a cognitive psychologist at University College London*. Click on the image to listen to the audio

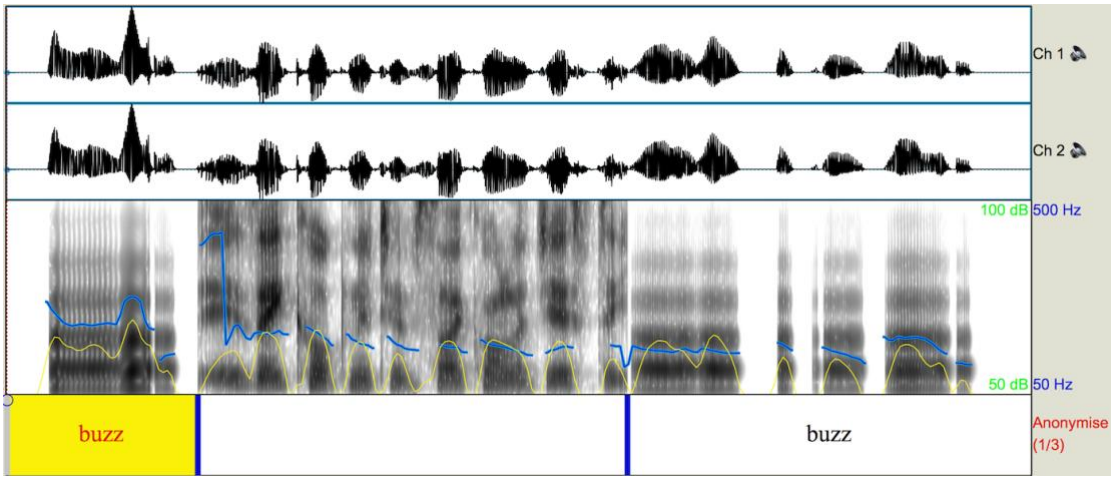

Figure 4: The manipulated speech signal, pitch contour and intensity profile of the utterance *Jenni Rodd is a cognitive psychologist at University College London*. Click on the image to listen to the audio[22]

[21] Note that since the recording, a podcast discussion, was obtained from the public domain, it has not been anonymised in the corpus.

[22] The audio snippets corresponding to the figures are also available at https://projekt.ht.lu.se/llc2/anonymisation.

In total, we anonymised approximately 1,300 personal pieces of information in LLC-2. The timestamps in the transcripts (see Section 3.2 above) helped us locate the information in the recordings with much less effort than if the transcripts had not been aligned with the recordings. This said, the manual nature of the task requires that corpus developers allow for a sufficient amount of time for completing it, which may prove impractical for larger corpora. However, the end result is worth the effort because it gives us a corpus that meets the ethical requirements of anonymity, which is mandatory for the public release of the audio material, and it also facilitates prosodic analyses on the corpus.

3.4. Applications of LLC-2 audio material

After tackling the challenges above, the LLC-2 audio material can be released to the public. The audio recordings are useful in a variety of areas in linguistics that, traditionally, have been outside the main focus of corpus linguistics. This section illustrates how the LLC-2 audio material can be used for investigations of the prosodic and temporal aspects of spoken interaction. It demonstrates three studies (Põldvere and Paradis 2019, 2020; Põldvere *et al*. submitted) based on data from LLC-2 that combined the orthographic transcriptions with instrumental analyses of the recordings to facilitate more thorough and, at times, even more reliable analyses of the phenomena in question.

Põldvere and Paradis (2019, 2020) were both concerned with a construction that previously had not received any attention in the literature, namely the reactive *what-x* construction. While Põldvere and Paradis (2020) set out to describe and define the constructional properties of the construction in LLC-2, Põldvere and Paradis (2019) tracked the development of the construction from LLC-1 to LLC-2, that is, over the past half a century. The LLC-1 audio material was made available to us by the *Survey of English Usage*. The analyses showed that the reactive *what-x* construction is a conventionalised construction in English that is characterised by a range of formal and functional properties that distinguish it from other, better-known *what*-constructions. One of these properties is prosody. Consider the utterance in bold in (1), which is an example of the reactive *what-x* construction in LLC-2.[23]

[23] Note that the transcriptions in this section have been slightly simplified in order to facilitate the task of the reader.

(1) <S051> I know it's ridiculous to plan Christmas already <pause/> although I did see <pause/> Christmas food in Sainsbury's yesterday

<S052> **what mince pies** <pause/>

<S051> all sorts of stuff

According to Põldvere and Paradis (2019, 2020), the reactive *what-x* construction always comprises the interrogative *what* and a subsequent complement, and its discursive meaning is to react to an immediately preceding turn to call it into question. In (1), *what* is followed by the noun phrase *mince pies*, used to react to the interlocutor's prior turn and to verify the specific Christmas food sold at Sainsbury's. However, an important property of the reactive *what-x* construction that cannot be derived from the orthographic transcription is that *what* always forms one and the same tone unit with the complement. This was determined in the studies through instrumental analyses of the construction in *Praat*.[24] Figure 5 illustrates the pitch contour of the reactive *what-x* construction in (1). As can be seen in the figure, *what* and *mince pies* form one and the same tone unit where *what* is realised as an unaccented pre-head of the unit, and the nuclear pitch accent, rise-fall, is on *pies*.[25] This information would have remained hidden to us had we not consulted the LLC-2 audio material.

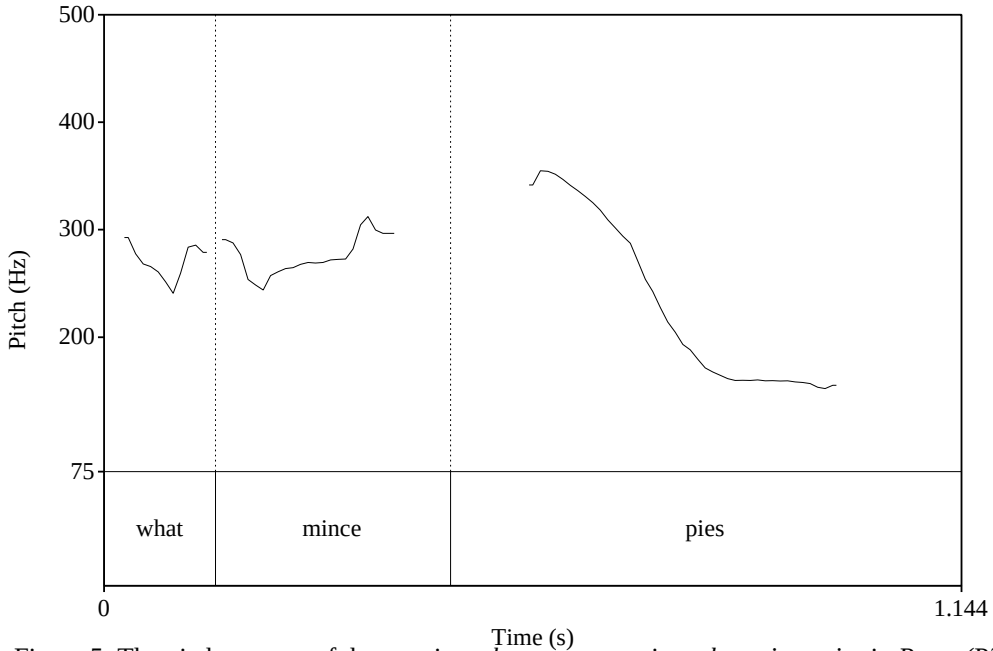

Figure 5: The pitch contour of the reactive *what-x* construction *what mince pies* in *Praat* (Põldvere and Paradis 2020: 320)

[24] In a few rare cases, the quality of the audio recordings was not good enough for instrumental analyses. In such cases, the recordings were auditorily inspected by both co-authors, and the decision as to the boundaries of the tone units and the types of nuclear pitch accents were made together.

[25] The prosodic analyses in Põldvere and Paradis (2019, 2020) follow the British tradition of intonation analysis where the basic unit is the tone, and where the direction of the pitch contour is a fall, rise, level, fall-rise or rise-fall (see, e.g., Cruttenden 1997).

Furthermore, the original audio recordings in the corpora helped us distinguish between the reactive *what-x* construction and a closely related *what*-construction, the pragmatic marker *what* (e.g., Brinton 2017). In many cases, the only property that sets the two constructions apart is that the pragmatic marker *what* always forms its own tone unit (e.g., *wh/\at # a b/\ird*),[26] which contributes to its interpretation as an expression of surprise and incredulity rather than a request for verification. Thus, the pragmatic marker *what* and the reactive *what-x* construction are two different constructions in English with distinct formal and functional characteristics. Without consulting the LLC-2 audio material, we would have missed this difference. In fact, this was a problem that we encountered in Põldvere and Paradis (2019), which included an additional analysis of the reactive *what-x* construction in Spoken BNC1994. Specifically, the missing audio data in the corpus meant that we were unable to classify eight per cent of the *what*-constructions included in the analysis. Furthermore, a comparison of the instrumental analysis of the LLC-1 audio material and the prosodic annotations revealed that not all instances of *what* in the transcripts had been assigned the correct prosodic pattern; in other words, what looked like the pragmatic marker *what* was in fact the reactive *what-x* construction, and vice versa. Thus, access to the LLC-1 audio material allowed us to validate the prosodic annotations against instrumental analyses and obtain more reliable results.

In Põldvere *et al*. (submitted), we used the LLC-2 audio material to investigate the timing of turns in conversational sequences where the speakers reproduce constructions from prior turns, called 'dialogic resonance' (Du Bois 2014). Consider the sequence in (2), taken from LLC-2, where the resonance is achieved through the speakers' choice of words and structures.

(2) <S002> yeah well so don't end up at home every day
<S003> I won't be at home every day <anon>Sara</anon>

According to Du Bois (2014), dialogic resonance emerges because speakers want to engage with the words of their interlocutors for various socio-communicative purposes. For example, previous work has showed that resonance is a fruitful way to express disagreement in spoken interaction (e.g., Dori-Hacohen 2017), as illustrated in (2). While Du Bois acknowledges the role of priming in resonance, this is not tested in his

[26] The hash sign (#) indicates a tone unit boundary between *what* and *a bird*, and /\ indicates a rising-falling pitch contour.

work. Instead, priming is the central mechanism of Garrod and Pickering's (2004) interactive alignment theory, which states that prior expression primes the reuse of the same linguistic representations by the next speaker. Thus, priming has a facilitating effect in resonance due to cognitive activation in the prior turn. In order to investigate the role of cognitive facilitation in resonance, we operationalised it as the time it takes for speakers to respond to the interlocutor's prior turn, based on the assumption that the timing of turns in conversation reflects the degree to which linguistic constructions are activated and accessible to the next speaker. The prediction was that transitions between speaker turns are faster in resonating sequences compared to when the turns are constructed from scratch. The results confirmed this prediction, showing that cognitive facilitation gives speakers the necessary tools to counter the temporal challenges of spontaneous conversation.

The analysis in Põldvere *et al*. (submitted) would not have been possible without the LLC-2 audio material. This is because the transcriptions in LLC-2 contain only limited information about turn transitions, showing whether a transition is a gap or an overlap but not its length in milliseconds. However, this information is crucial for systematic investigations of the timing of turns in conversation. In order to extract reliable measurements of turn transitions in the data, we used the multimodal annotation tool ELAN. The advantage of using ELAN over other speech analysis software such as *Praat* is that ELAN allows for the annotation of the speech signal using multiple tiers that can be created freely by the analyst. Moreover, the length of the annotations in milliseconds can be easily exported to a spreadsheet or database software for statistical analysis. Figure 6 illustrates the speech signal and the corresponding annotation of the conversational sequence in (2) above. As can be seen in the figure, the annotation scheme includes the orthographic transcription of the utterances in the conversational sequence, and the type of transition between the utterances, in this case a gap. The exported data reveal that the length of the gap is eight milliseconds, which is very fast considering that the dialogic function of the response is to express disagreement, a dispreferred response. The rest of the annotations in Figure 6 need not concern us here.

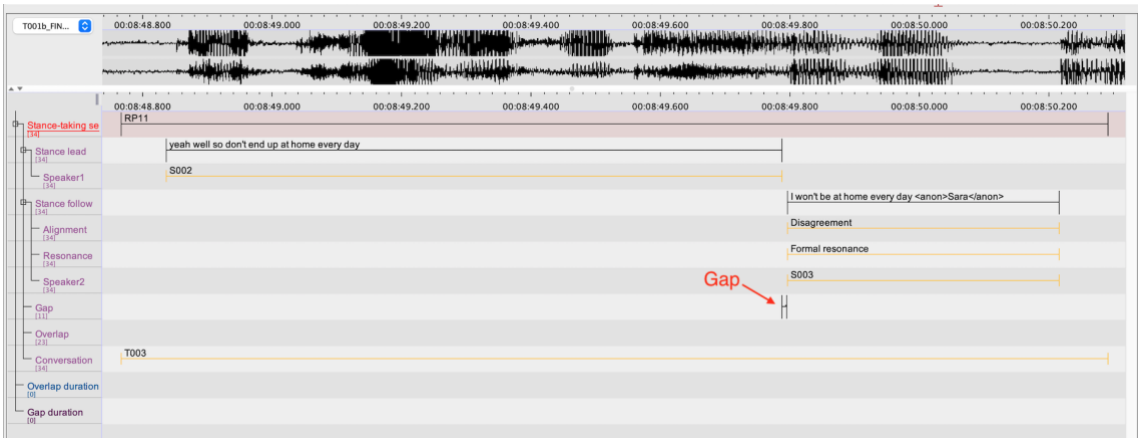

Figure 6: The illustration of a gap between the resonating utterances expressing disagreement, *yeah well so don't end up at home every day* and *I won't be at home every day <anon>Sara</anon>* in ELAN

In sum, the studies above show that, with access to the LLC-2 audio material, and the appropriate software, users have at their disposal all the necessary tools to carry out thorough and reliable analyses of prosody and turn-taking in spoken interaction, and therefore promote the extension of corpus linguistics in new directions. The cost and effort associated with overcoming the methodological challenges of preparing the audio material for public release has been a small price to pay for such a gain.

4. Conclusion and future work

The aim of this article has been to describe key challenges of preparing and releasing audio material for spoken data and to propose solutions to these challenges. We have focused on two challenges that we had to tackle during the compilation of LLC-2: 1) the alignment of the orthographic transcriptions with the audio files and 2) the anonymisation of personal information in the recordings. Audio-to-text alignment was necessary because it allows users to easily link relevant sections in the transcripts to the corresponding locations in the audio files. We opted for a solution that involved inserting timestamps by means of *InqScribe* in front of speaker turns to indicate to the users where each turn begins. As shown, this solution can be effectively combined with more sophisticated automatic segmentation techniques (e.g., the WebMAUS forced alignment system). The second challenge concerned the anonymisation of personal information in the audio recordings, which was mandatory in order to abide by the ethical and legal principles of privacy and data protection. For the best result possible, we used a *Praat* script developed by Hirst (2013). The script replaces all personal information in the recordings with a sound that makes the lexical information

incomprehensible but retains the prosodic characteristics of the original speech signal. The advantage of this technique over some of the other techniques suggested in the literature is that it is reliable and makes possible a wide variety of linguistic analyses, including prosody.

The release of the LLC-2 audio material together with the transcripts is unique because it opens up research opportunities that extend the scope of corpus linguistics in new and exciting directions. This article has focused on two areas that are conspicuously under-researched in spoken corpus research: prosody and turn-taking. Drawing on three studies based on data from LLC-2, we have demonstrated that the LLC-2 audio material can be used to perform thorough and reliable investigations of the prosodic and temporal aspects of spoken interaction using freely available speech analysis and annotation tools. In our view, the opportunities that the LLC-2 audio recordings offer for spoken corpus research overweigh the methodological challenges of making them publicly available. Therefore, future corpus developers are encouraged to factor in the time and effort of tackling these challenges. At the same time, we acknowledge that the techniques presented here may be more suitable for smaller-scale corpora such as LLC-2 rather than larger, multi-million-word national corpora. This is mainly due to the considerable amount of manual effort needed, particularly in the annotation of personal pieces of information in *Praat*. This said, the rapid technological advances in machine learning and audio-to-text technologies give us hope that, in the not-too-distant future, these techniques can be scaled up to larger corpora, too. In the meantime, the present techniques could be applied to a subset of a larger corpus in order to facilitate prosodic and temporal analyses on, at least, a part of it.

Future work on LLC-2 involves making the recordings and transcripts available from the free corpus management and analysis system *Corpuscle* (Meurer 2012). *Corpuscle* will enable the implementation of various corpus linguistic techniques on LLC-2, and the possibility to carry out restricted searches on the corpus data based on the many demographic categories available in the metadata. The release of LLC-2 from *Corpuscle* also means that users will no longer have to navigate the individual XML transcription files and WAV audio files to be able to listen to relevant sections of the transcripts. Instead, this process will be made considerably quicker by the audio playback function of *Corpuscle* in which case a click on the transcription immediately plays back the corresponding part of the recording. The most promising feature of

Corpuscle for LLC-2 is that the audio playback works on a turn-by-turn basis, meaning that the timestamps in the transcripts will be sufficient for setting it up. We hope that the combination of downloadable and time-aligned transcription and audio files with online audio snippets will lead to even more diverse uses of LLC-2 and facilitate seamless experiences of using the corpus.

REFERENCES

Aijmer, Karin. 1996. *Conversational Routines in English: Convention and Creativity*. London: Longman.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word combinations. In Anthony P. Cowie ed. *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press, 101–122.

Andersen, Gisle. 2016. Semi-lexical features in corpus transcription: Consistency, comparability, standardisation. *International Journal of Corpus Linguistics* 21/3: 323–347.

Atkins, Sue, Jeremy Clear and Nicholas Ostler. 1992. Corpus design criteria. *Literary and Linguistic Computing* 7/1: 1–16.

Baghai-Ravay, Ladan, Greg Kochanski and John Coleman. 2009. Precision of phoneme boundaries derived using Hidden Markov Models. *Proceedings of INTERSPEECH 2009, Tenth Annual Conference of the Interantional Speech Communication Association*, 2879–2882.

Boersma, Paul. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5/9–10: 341–345.

BNC Consortium. 2007. *The British National Corpus*, version 3 (BNC XML Edition). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk. (9 April, 2021.)

Brinton, Laurel J. 2017. *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge: Cambridge University Press.

Calhoun, Sasha, Jean Carletta, Jason M. Brenier, Neil Mayo, Dan Jurafsky, Mark Steedman and David Beaver. 2010. The NXT-format Switchboard Corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation* 44/4: 387–419.

Clayman, Steven E. 2002. Sequence and solidarity. In Shane R. Thye and Edward J. Lawler eds. *Group Cohesion, Trust and Solidarity*. Oxford: Elsevier, 229–253.

Coleman, John, Ladan Baghai-Ravary, John Pybus and Sergio Grau. 2012. Audio BNC: The Audio Edition of the Spoken British National Corpus. http://www.phon.ox.ac.uk/AudioBNC (24 February, 2020.)

Cosi, Piero, Daniele Falavigna and Maurizio Omologo. 1991. A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *Proceedings of EUROSPEECH 1991, Second European Conference on Speech Communication and Technology*, 693–696.

Crowdy, Steve. 1994. Spoken corpus transcription. *Literary and Linguistic Computing* 9/1: 25–28.

Cruttenden, Alan. 1997. *Intonation*. Cambridge: Cambridge University Press.

Diemer, Stefan, Marie-Louise Brunner and Selina Schmidt. 2016. Compiling computer-mediated spoken language corpora: Key issues and recommendations. *International Journal of Corpus Linguistics* 21/3: 348–371.

Dori-Hacohen, Gonen. 2017. Creative resonance and misalignment stance: Achieving distance in one Hebrew interaction. *Functions of Language* 24/1: 16–40.

Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1/1: 71–106.

Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25/3: 359–410.

Edwards, Jane A. 1995. Principles and alternative systems in the transcription, coding, and mark-up of spoken discourse. In Geoffrey Leech, Greg Myers and Jenny Thomas eds. *Spoken English on Computer: Transcription, Mark-Up and Application*. New York: Longman, 19–34.

Garrod, Simon and Martin J. Pickering. 2004. Why is conversation so easy? *TRENDS in Cognitive Sciences* 8/1: 8–11.

Gerstenberg, Annette, Valerie Hekkel, Julie Marie Kairet and Adélie Soumier-Vendé. 2017. *LangAge Corpora: Resources for Language and Aging Research*. Poster presentation at CLARe3, Berlin, Germany.

Grabe, Esther. 2004. Intonational variation in urban dialects of English spoken in the British Isles. In Peter Gilles and Jörg Peters eds. *Regional Variation in Intonation*. Tübingen: Niemeyer, 9–31.

Greenbaum, Sidney and Jan Svartvik. 1990. The *London-Lund Corpus of Spoken English*. In Jan Svartvik ed. *The London-Lund Corpus of Spoken English: Description and Research*. Lund: Lund University Press, 11–59.

Gries, Stefan Th. and Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide and James Pustejovsky eds. *Handbook of Linguistic Annotation*. Berlin: Springer, 379–409.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103.

Hasund, Ingrid Kristine. 1998. Protecting the innocent: The issue of informants' anonymity in the COLT corpus. In Antoinette Renouf ed. *Explorations in Corpus Linguistics*. Amsterdam: Rodopi, 13–28.

Hirst, Daniel. 2013. Anonymising long sounds for prosodic research. In Brigitte Bigi and Daniel Hirst eds. *Tools and Resources for the Analysis of Speech Prosody*. Aix-en-Provence: Laboratoire Parole et Langage, 36–37.

Hoey, Elliott M. and Kobin H. Kendrick. 2017. Conversation Analysis. In Annette M. B. de Groot and Peter Hagoort eds. *Research Methods in Psycholinguistics and the Neurobiology of Language: A Practical Guide*. Hoboken: Wiley-Blackwell, 151–173.

Hoffmann, Sebastian and Sabine Arndt-Lappe. Submitted. Better data for more researchers – Using the audio features of BNCweb.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith, David Lee and Ylva Berglund Prytz. 2008. *Corpus Linguistics with BNCweb – A Practical Guide*. Frankfurt am Main: Peter Lang.

Hosom, John-Paul. 2000. *Automatic Time Alignment of Phonemes Using Acoustic-Phonetic Information*. Hillsboro, OR: Oregon Health and Science University dissertation.

Hosom, John-Paul. 2009. Speaker-independent phoneme alignment using transition-dependent states. *Speech Communication* 51/4: 352–368.

InqScribe. 2005–2020. Computer software. https://www.inqscribe.com/ (9 April, 2021.)

Kaufmann, Anita. 2002. Negation and prosody in British English. *Journal of Pragmatics* 34/10: 1473–1494.

Kendrick, Kobin H. and Francisco Torreira. 2015. The timing and construction of preference: A quantitative study. *Discourse Processes* 52/4: 255–289.

Kimps, Ditte. 2018. *Tag Questions in Conversation: A Typology of their Interactional and Stance Meanings*. Amsterdam: John Benjamins.

Kirk, John M. 2016. The Pragmatic Annotation Scheme of the *SPICE-Ireland Corpus*. *International Journal of Corpus Linguistics* 21/3: 299–322.

Kirk, John M. and Gisle Andersen. 2016. Compilation, transcription, markup and annotation of spoken corpora. *International Journal of Corpus Linguistics* 21/3: 291–298.

Kisler, Thomas, Uwe Reichel and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech & Language* 45: 326–347.

Leech, Geoffrey. 2004. Adding linguistic annotation. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. http://users.ox.ac.uk/~martinw/dlc/chapter2.htm (24 February, 2020.)

Lenk, Uta. 1998. *Marking Discourse Coherence: Functions of Discourse Markers in Spoken English*. Tübingen: Gunter Narr Verlag.

Lin, Phoebe. 2018. *The Prosody of Formulaic Sequences: A Corpus and Discourse Approach*. London: Bloomsbury.

Love, Robbie, Claire Dembry, Andrew Hardie, Vaclav Brezina and Tony McEnery. 2017. The Spoken BNC2014: Designing and building a corpus of everyday conversations. *International Journal of Corpus Linguistics* 22/3: 319–344.

McEnery, Tony. 2018. The Spoken BNC2014: The corpus linguistic perspective. In Vaclav Brezina, Robbie Love and Karin Aijmer eds. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge, 10–15.

Meurer, Paul. 2012. *Corpuscle* – A new corpus management platform for annotated corpora. In Gisle Andersen ed. *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian*. Amsterdam: John Benjamins, 29–50.

Nelson, Gerald. 2002. *Markup Manual for Spoken Texts.* http://ice-corpora.net/ice/index.html (24 February, 2020.)

Nelson, Gerald, Sean Wallis and Bas Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.

Ochs, Elinor. 1979. Transcription as theory. In Elinor Ochs and Bambi B. Schiefflen eds. *Developmental Pragmatics*. New York: Academic Press, 43–72.

Oostdijk, Nelleke and Lou Boves. 2008. Preprocessing speech corpora: Transcription and phonological annotation. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* Vol. 1. Berlin: Mouton de Gruyter, 642–663.

Paradis, Carita. 1997. *Degree Modifiers of Adjectives in Spoken British English*. Lund: Lund University Press.

Paradis, Carita. 2008. Configurations, construals and change: Expressions of degree. *English Language and Linguistics* 12/2: 317–343.

Põldvere, Nele and Carita Paradis. 2019. Motivations and mechanisms for the development of the reactive *what-x* construction in spoken dialogue. *Journal of Pragmatics* 143: 65–84.

Põldvere, Nele and Carita Paradis. 2020. 'What and then a little robot brings it to you?' The reactive *what-x* construction in spoken dialogue. *English Language and Linguistics* 24/2: 307–332.

Põldvere, Nele, Matteo Fuoli and Carita Paradis. 2016. A study of dialogic expansion and contraction in spoken discourse using corpus and experimental techniques. *Corpora* 11/2: 191–225.

Põldvere, Nele, Victoria Johansson and Carita Paradis. In press a. *A Guide to the London-Lund Corpus 2 of Spoken British English*. Lund Studies in English. Lund: Centre for Languages and Literature, Lund University.

Põldvere, Nele, Victoria Johansson and Carita Paradis. In press b. On the *London-Lund Corpus 2*: Design, challenges and innovations. *English Language and Linguistics* 25/3.

Põldvere, Nele, Victoria Johansson and Carita Paradis. Submitted. Resonance in dialogue: The interplay between intersubjective motivations and cognitive facilitation.

Roberts, Felicia, Alexander L. Francis and Melanie Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication* 48/9: 1079–1093.

Roberts, Seán G., Francisco Torreira and Stephen C. Levinson. 2015. The effects of processing and sequence organization on the timing of turn taking: A corpus study. *Frontiers in Psychology* 6: 1–16.

Romero-Trillo, Jesús. 2014. 'Pragmatic punting' and prosody. In María de los Ángeles Gómez González, Francisco José Ruiz de Mendoza Ibáñez, Francisco Gonzálvez-García and Angela Downing eds. *The Functional Perspective on Language and Discourse: Applications and Implications*. Amsterdam: John Benjamins, 209–222.

Sauer, Simon and Anke Lüdeling. 2016. Flexible multi-layer spoken dialogue corpora. *International Journal of Corpus Linguistics* 21/3: 419–438.

Schiel, Florian. 1999. Automatic phonetic transcription of non prompted speech. In John J. Ohala, Yoko Hasegawa, Manjari Ohala, Daniel Granville and Ashlee C. Baile eds. *Proceedings of ICPhS 1999, Fourteenth International Congress of Phonetic Sciences*, 607–610.

Schmidt, Thomas. 2016. Good practices in the compilation of FOLK, the *Research and Teaching Corpus of Spoken German*. *International Journal of Corpus Linguistics* 21/3: 396–418.

Stenström, Anna-Brita. 1984. *Questions and Responses in English Conversation*. Malmö: Gleerup.

Stenström, Anna-Brita, Gisele Andersen, Kristine Hasund, Kristina Monstad and Hanne Aas. 1998. *User's Manual to Accompany The Bergen Corpus of London Teenage Language* (COLT). Bergen: Department of English, University of Bergen.

Svartvik, Jan and Randolph Quirk eds. 1980. *A Corpus of English Conversation*. Lund: Gleerup.

Thompson, Paul. 2004. Spoken language corpora. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. http://users.ox.ac.uk/~martinw/dlc/chapter5.htm (9 April, 2021.)

UCL Survey of English Usage. 2020. https://www.ucl.ac.uk/english-usage/ (5 April, 2021.)

Wallis, Sean, Gerald Nelson and Bas Aarts eds. 2006. *The British Component of the International Corpus of English* (ICE-GB), Release 2. London: *Survey of English Usage* computer software.

Weisser, Martin. 2017. Annotating the ICE corpora pragmatically – Preliminary issues & steps. *ICAME Journal* 41/1: 181–214.

Wichmann, Anne. 2008. Speech corpora and spoken corpora. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* Vol. 1. Berlin: Mouton de Gruyter, 187–206.

Wichmann, Anne. 2011. Grammaticalization and prosody. In Bernd Heine and Heiko Narrog eds. *The Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press, 331–341.

Wichmann, Anne, Anne-Marie Simon-Vandenbergen and Karin Aijmer. 2010. How prosody reflects semantic change: A synchronic case study of *of course*. In Kristin Davidse, Lieven Vandelanotte and Hubert Cuyckens eds. *Subjectification, Intersubjectification and Grammaticalization*. Berlin: Mouton de Gruyter, 103–154.

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 LREC 2006, Fifth International Conference on Language Resources and Evaluation*, 1556–1559.

Yuan, Jiahong, Wei Lai, Chris Cieri and Mark Liberman. 2018. Using forced alignment for phonetics research. In Chu-Ren Huang, Peng Jin and Shu-Kai Hsieh eds. *Chinese Language Resources and Processing: Text, Speech and Language Technology*. Springer.

Corresponding author
Nele Põldvere
Centre for Languages and Literature
Lund University
Box 201
221 00 Lund
Sweden
Email: nele.poldvere@englund.lu.se

received: January 2020
accepted: March 2021

RiCL Research in Corpus Linguistics

Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription

Marie-Louise Brunner ‑ Stefan Diemer
Trier University of Applied Sciences / Germany

Abstract – The article discusses how to integrate annotation for nonverbal elements (NVE) from multimodal raw data as part of a standardized corpus transcription. We argue that it is essential to include multimodal elements when investigating conversational data, and that in order to integrate these elements, a structured approach to complex multimodal data is needed. We discuss how to formulate a structured corpus-suitable standard syntax and taxonomy for nonverbal features such as gesture, facial expressions, and physical stance, and how to integrate it in a corpus. Using corpus examples, the article describes the development of a robust annotation system for spoken language in the corpus of *Video-mediated English as a Lingua Franca Conversations* (ViMELF 2018) and illustrates how the system can be used for the study of spoken discourse. The system takes into account previous research on multimodality, transcribes salient nonverbal features in a concise manner, and uses a standard syntax. While such an approach introduces a degree of subjectivity through the criteria of salience and conciseness, the system also offers considerable advantages: it is versatile and adaptable, flexible enough to work with a wide range of multimodal data, and it allows both quantitative and qualitative research on the pragmatics of interaction.

Keywords – corpus annotation; corpus transcription; multimodality; nonverbal elements; spoken discourse; video-mediated communication; gestures

1. Multimodality as part of rich data: The transcriber's dilemma

Complex or 'rich' data poses specific problems in terms of corpus integration. Paralinguistic elements, such as prosody, overlap, laughter in audio data, or nonverbal elements such as gaze, gestures and background interaction in video data, are introducing a level of complexity that is difficult to integrate as part of a replicable and structured transcription system. The question of how to handle such rich data has become increasingly urgent, as more and more datasets have become available through online sources or multimodal compilation projects (cf. Brunner *et al*. 2017). Research acknowledges multimodality as an integral part of the meaning-making process, and studies on multimodal discourse, such as Kress (2011) and Scollon and LeVine (2004),

have established a comprehensive view on language in use as "always and inevitably constructed across multiple modes of communication, including speech and gesture […]." (Scollon and LeVine 2004: 1f.). The realization that the lexical level is only one of many modes and thus only a partial means of meaning making (cf. Kress 2011: 46) creates a problem for corpus researchers. Bezemer and Jewitt (2010: 194) caution that "[m]ultimodality is an eclectic approach" and argue that researchers are faced with a dilemma:

> Too much attention to many different modes may take away from understanding the meanings of a particular mode; too much attention to one single mode and one runs the risk of 'tying things down' to just one of the many ways in which people make meaning. (Bezemer and Jewitt 2010: 194)

A possible solution is to ensure that both corpus data and corpus architecture allow the integration of additional modes, creating the possibility to study various features either independently or in correlation. Multimodality is, of course, a very broad term and, while we consider key features such as paralanguage (e.g. laughter) as part of the general multimodal setting (and as necessary component of spoken corpus transcription), in this article we will focus on the representation of what we term 'nonverbal elements' (NVE), comprising gestures, facial expressions, gaze, and physical stance, as well as camera shifts and background events. This use of NVE constitutes a slight expansion of Adolphs and Carter's term 'nonverbal features' for "gestures which exist in and complement spoken discourse" (Adolphs and Carter 2013: 145), as we also include some affordances of the medium that serve a similar purpose, such as camera shifts and visual and auditory background events (e.g. a person being visible in the background, intruding, or talking to one of the speakers). As part of the transcription and annotation process, corpus compilers and annotators have to achieve the balancing act between preserving and documenting nonverbal features as far as possible in the transcriptions and focusing on those features that are salient in the discourse context. Salience here refers to NVE contributing to or supporting meaning making, as well as NVE that are referred to on a verbal level or that refer to something that is discussed in the conversation (see also Section 3.1). This is, of course, difficult, and researchers have to choose how much of the rich information in a multimodal dataset can and should be included in the finished corpus, either as part of the transcription or as one of various corpus components. This problem of choice is compounded by the complex nature of the data. This rich data creates a second dilemma:

finding a standard way of transcribing it. As DuBois (1991: 73) points out, “there is not, nor ever can be, a single standard way of putting spoken word to paper.”

The question we will explore in this article is how to integrate annotation for NVE elements from multimodal raw data as part of a standard lexical transcription corpus. We argue that it is essential to include multimodal elements when investigating conversational data wherever possible, and that in order to integrate these elements, a structured approach to the complex, unstructured multimodal data is needed. Artificial Intelligence-supported automated gesture recognition does not (yet) provide a satisfactory solution here, and the complex nature of multimodality makes manual annotation necessary in order to obtain gold standard corpus data. We thus need a standard syntax and taxonomy for manually annotating nonverbal features.

We present our approach to creating and implementing such a standard system in the corpus of *Video-mediated English as a Lingua Franca Conversations* (ViMELF 2018). Using examples from the corpus, we describe the bottom-up development of a manual annotation system for spoken language that takes into account previous research on multimodal features, focuses on salience and simplification, and uses a standard syntax. We will also illustrate its potential use in discourse research. Our aim is the creation of a concise and robust transcription system which can be used with a large variety of search tools by researchers from various disciplines who do not need any previous knowledge in gesture research in order to read and understand the data. We see possible uses in varied fields, such as corpus-based multimodal discourse analysis, corpus linguistics, conversation analysis, interactional (socio)linguistics, World Englishes, English as a Lingua franca, and language acquisition.

2. Existing approaches to the description and transcription of nonverbal elements

Multimodal features, and in particular nonverbal elements such as gestures, pose a considerable problem for corpus compilation. The crucial role of gestures in interaction is frequently underlined (e.g. Kendon 2004; Goodwin and Goodwin 2000), and gestures have been studied extensively in multiple branches of linguistics, such as conversation analysis, language acquisition, cognitive linguistics, psycholinguistics, forensic linguistics, multimodal discourse analysis, linguistic anthropology, as well as in psychology. There have also been repeated calls to integrate multimodal data as part of

corpus data (e.g. Adolphs and Carter 2013). However, there is, to our knowledge, no generally recognized and practical transcription system that manages to capture this complex dynamic interaction between gesture, context, and talk. The main problem to overcome in developing an annotation system for nonverbal elements is their complex nature in terms of contribution to discourse, which gesture research has variously commented upon. In our analysis of the various approaches, we will distinguish 'describing' from 'transcribing' nonverbal elements.

2.1. Describing nonverbal elements

Adam Kendon, one of the foremost gesture researchers, describes gestures as utterances that contribute to human understanding like vocal elements, as visual behavior with a communicative and not only informative or expressive function (cf. Kendon 2004). From a psycholinguistic perspective, David McNeill and Duncan call gestures dimensions of social interaction that "open a 'window' onto thinking" (McNeill and Duncan 2000: 143). In his own gesture research, Jürgen Streeck foregrounds their complexity as "largely improvised, heterogeneous, partly conventional, partly idiosyncratic, partly culture-specific, partly universal practice to produce situated understandings" (Streeck 2009: 5). For Charles Goodwin, the acknowledged expert on embodied talk in interaction, all interaction is embodied interaction, movement requires talk and talk requires gestures, and all three create a whole that is different from and greater than the individual parts, as "each individual sign is partial and incomplete" (Goodwin 2007: 199). These descriptions set the scene for the various entailed research perspectives.

Gestures can be described structurally, that is, which body parts are involved, the positioning of these body parts, and movement phases, and how this correlates with prosody or speech in general (e.g. Kendon 1980; McNeill 1992). Another way of describing gestures is by describing the semiotic and semantic content of the gesture in combination with underlying cognitive processes, for example, iconic, metaphorical, indexical, or beat gestures (e.g. Kendon 2004; McNeill 2008; Calbris 2011). A holistic approach to describing nonverbal behavior incorporates gesture as part of a broader concept of embodied action, showing nonverbal elements, the body and its positioning with respect to others and the environment, objects and the surroundings, as well as activities that are being carried out in addition to and in interaction with the verbal level

(e.g. Goodwin 2000; Mondada 2014). These general perspectives on describing gestures are variously employed and adapted by the respective linguistic disciplines.

2.2. Transcribing nonverbal elements

Conversation analysis (CA) researchers routinely include prosodic, paralinguistic, and nonverbal elements in their transcriptions, and there are established annotation systems for "the delivery of talk and other bodily conduct" (Hepburn and Bolden 2013: 57) going back to the Jeffersonian annotation scheme (e.g. Jefferson 1973; Sacks *et al*. 1978). The CA scheme is "a shared, standard system for rendering talk-in-interaction" (Hepburn and Bolden 2013: 75) which is insightful, detailed, and highly relevant for studying spoken discourse. Its main shortcoming from the perspective of corpus linguistics is its limited suitability for quantitative research. CA transcription has been characterized as somewhat unsystematic in its representation of selected features (e.g. DuBois 1991). It is also highly individualized depending on the transcribers' research focus and does not provide a general framework for multimodality, but rather allows the inclusion of selected multimodal features as needed when transcribing the data for a particular purpose of analysis. Researchers in the fields of interactional sociolinguistics, semiotics, and pragmatics have also been studying and transcribing nonverbal behavior in discourse. Gestures are transcribed variously as part of a multi-layered score (Kendon 2004), as aligned descriptions with accompanying illustrations (Streeck 2009), as series of images illustrating stages and aligned with the text (Mondada 2014), as dynamic comic-like transcript inserts, or as a combination of all of the above (McNeill 2008, 2017).

For various reasons, these transcription schemes are not ideal for a corpus context: approaches that strive to be descriptive tend to become increasingly elaborate and difficult to understand. Examples are McNeill's verbal transcriptions (McNeill 2008, 2017) or the complex 'Linguistic Annotation System for Gestures' (LASG) developed by Bressem *et al.* (2013). Approaches that classify gestures through additional visual elements (e.g. Mondada 2014) are difficult to analyze quantitatively, and approaches that focus on interaction dynamics (e.g. Goodwin 2007) introduce a considerable degree of interpretation. With the rapid development of Artificial Intelligence (AI) supported automatic gesture recognition since 2015, attempts to automatically map and systematize gestures as part of multimodal construction grammar are under way (e.g. Joo *et al*. 2017). Though this research direction looks promising with further advances in image

recognition, results so far are limited to a basic physical and very detailed taxonomy of hand gestures and body orientation in TV news data based on gold standard, manually transcribed corpora.

When compiling ViMELF, the approaches described above were considered unsuitable for the purpose of providing a manual annotation system for nonverbal elements that is sufficiently systematized, yet robust, and accounts for all salient features in an interactional context. This prompted the development of the transcription system which is described in Sections 3 and 4.

3. TRANSCRIBING VIMELF

3.1. Data and general transcription guidelines

ViMELF (2018) is a small corpus of 20 dyadic video-mediated conversations in an informal setting between previously unacquainted participants from Germany, Spain, Italy, Finland and Bulgaria, using English as a Lingua Franca. The corpus comprises 113,677 words in the plain text version and 154,472 tokens including annotation (NVE, paralanguage, and affordances of the medium). The gestural annotation is integrated as part of the general transcript rather than creating a separate layer for nonverbal elements (see also Section 4.2). There are 7,449 NVEs in total, which are distributed over 6,463 instances of transcribed nonverbal behavior. The full corpus length amounts to 744.5 minutes (ca. 12.5 hours) of recorded conversation with an average conversation length of 37.23 minutes. The corpus was published in 2018 by the research group of the *Corpus of Academic Spoken English* (CASE) at Trier University of Applied Sciences (Germany), where the corpus is also hosted.[1] It is freely available for research, including the anonymized audio and video recordings. The transcripts provide timestamps every 30 seconds as a simple alignment feature in order to facilitate retrieving the corresponding audio or video sequences for a more comprehensive analysis.[2]

ViMELF was transcribed and annotated manually by a team of more than 60 transcribers on the basis of Dressler and Kreuz's (2000) synthetic transcription conventions which were then extended for the particular conditions of spoken computer-mediated communication (CMC) in an international context. In developing an annotation

[1] For further information on ViMELF (2018), see the project website at http://umwelt-campus.de/case
[2] Timestamps were omitted in the transcribed examples used in this paper to facilitate reading.

system for this particular setting, the transcription team followed, as much as possible, DuBois' (1991) and Edwards' (1993) guidelines for spoken discourse transcription, which still constitute best practice in the field. DuBois' (1991) maxims for transcription are: (i) a clear definition of categories, (ii) accessibility, including the use of notations that maximizes access and are easily and intuitively readable, (iii) robustness, (iv) economy, and (v) adaptability. Edwards (1993) requires the established categories to be (i) discriminable, (ii) exhaustive, and (iii) contrastive, with the aim of creating a systematic and predictable scheme that allows multiple transcribers to work on the data while ensuring consistency and retrievability.

Because multimodal corpora are impossible to transcribe fully, both DuBois and Edwards recommend that transcribers have to be selective and select a finite number of features for transcription. The key criterion for choosing which features to transcribe is salience, in particular in relation to multimodal elements supplementing the lexical level. Our use of salience here refers to "a property of a linguistic item or feature that makes it in some way perceptually and cognitively prominent" (Kerswill and Williams 2002: 81), that is, contributing in some way to meaning making. As Norris (2002: 118) points out, "salience derives from the interaction," which means that multimodal elements can enhance the verbal level or even acquire their own salience independently of the lexical level (e.g. pointing). DuBois' robustness and economy maxims also reflect the need to establish salient categories, while Edwards' demand for an exhaustive set of categories is more difficult to maintain in this respect. Conversely, not all multimodal elements are salient; they can also be incidental or redundant. To determine salience with a maximum degree of objectivity, transcribers need to compare their respective perceptions during transcription and also consider the baseline of the respective dataset in order to produce a consistent corpus transcription. A speaker may have a certain base speaking speed which can be either slow or fast, so slow speaking in itself may not be salient. Deviating from the base speed may, however, foreground particular items and thus establish salience. The same is true of habitual gestures such as scratching one's nose, in comparison to one-time gestures that may convey meaning in this particular context. Scratching one's nose while pausing and saying *Ummmm* can, for example, convey skepticism. The criterion of salience introduces a certain unavoidable degree of subjective interpretation, but transcription would be impossibly detailed without it. Salience is the only way to satisfy Edwards' maxim for an exhaustive set of categories in a multimodal

dataset, and we thus consider salience to be the most important maxim when transcribing multimodal data.

In terms of transcription procedure, DuBois advocates a transcriber-centered system that allows the transcribers' increasing experience during the transcription process to filter back into the system:

> The system should be convenient and comfortable to use, reasonably easy to learn, and through its implicit categories it should promote insightful perception and classification of discourse phenomena, which in the end may feed back into advances in the system itself. [...] Through the experience of transcribing the transcriber is constantly learning about discourse. (DuBois 1991: 75)

DuBois also cautions that the system needs to be flexible:

> It is the transcriber, immersed in the recorded speech event and grounded in discourse theory, who is in a position to [...] advance the potential of the transcription system and its theoretical framework. (DuBois 1991: 75)

After describing the data and the general guidelines that were followed during transcription and annotation of ViMELF, the transcription process itself is presented in Section 3.2.

3.2. Transcription process

The ViMELF transcription process was designed to ensure that the guidelines presented in Section 3.1 were observed, and that transcribers had opportunities for feedback during the transcription process. The process consisted of a pilot transcription phase, followed by three consecutive main transcription phases: pilot transcription phrase, first transcription phase, and second transcription phase.

In the pilot transcription phase, senior project transcribers transcribed the same randomly selected conversations to identify key issues and potential inconsistencies. The transcripts were then compared, and the guidelines formulated and refined in several transcription rounds until inter-transcriber reliability was above 95 percent.

During the first transcription phase, 50 student transcribers were employed in three consecutive rounds as data became available. The student transcribers were trained in specific transcription tutorials that included a parallel transcription of corpus data by all

transcribers and a joint analysis and discussion of inconsistencies. After training, each student transcriber then transcribed at least 30 minutes of conversation; some conversations were transcribed by multiple student transcribers to check for remaining inconsistencies. Transcription was done with the help of the transcription software *F4trankript*, which facilitates close analysis of the audio and video data through features such as repetition looping, timestamping, and low-speed playback. Transcribers were free to either integrate verbal, nonverbal and paralinguistic features at the same time or to work on each feature (e.g. lexis, pauses, laughter, nonverbal elements) consecutively, as both techniques yielded data of comparable quality. Student transcribers were regularly polled on transcription issues. Based on the results of the transcriber polls, average duration for transcribing one minute of audiovisual data is around two hours for a novice and one hour for a senior transcriber. Not surprisingly, the areas where the most significant issues and inconsistencies were reported were the identification and transcription of nonverbal elements, the transcription of paralinguistic elements, in particular laughter, and the identification of intonation units. This prompted the development and further refinement of separate guidelines for the transcription of nonverbal elements as presented in this article. Separate guidelines were also created for the treatment of paralanguage, in particular laughter (for a discussion of ViMELF transcription guidelines for laughter see Brunner *et al.* 2017).

The second phase of the main transcription consisted of a thorough second transcription and correction by six senior project transcribers. The senior transcribers compared transcripts in regular meetings and discussed inconsistencies and general issues. Remaining inconsistencies were consolidated, and the guidelines were further elaborated if needed and then fed back into the next round of transcriptions. Inter-transcriber reliability at the end of this phase was evaluated at 98 percent.

The third phase consisted of a final correction by four project coordinators to ensure consistency of the final dataset. Project coordinators and senior project transcribers met regularly to discuss differences in transcription, issues of salience, and problematic features.

In sum, regular team meetings at all transcription stages ensured transcriber input on desirable adjustments in the transcription system, contributing to a data-driven, bottom-up formulation of guidelines. The resulting transcription guidelines for ViMELF contain provisions for:

(i) lexical transcription,

(ii) spoken language features (cut-offs, overlap, liaisons, latching),

(ii) prosody (intonation, pitch, volume, speed, pauses) and paralinguistic features (laughter, coughing, sighing, loud breathing),

(iv) nonverbal elements (gestures, facial expressions, gaze, physical stance, camera shifts, background events).

In addition, some specific ELF and video-mediated features are also transcribed, such as code-switching, non-standard pronunciations, and technical issues such as echo.[3] While the guidelines represent the result of an elaborate process, the availability of the anonymized raw data as part of the corpus specifically allows further development and inclusion of additional features at need, depending on the interest of future researchers and transcribers. In the context of this paper, we will focus on just one of the most challenging features to illustrate the design and compilation of transcription guidelines: the transcription of nonverbal behavior.

4. Developing a transcription system for nonverbal elements

4.1. Nonverbal elements in interaction: Examples from ViMELF

In his seminal 1991 article on transcribing spoken discourse data, DuBois does not provide for a multimodal transcription system, but already indicates the need for further research in that direction:

> There are several dimensions along which further development can be hoped for in the coming years ─for example [...] nonverbal cues like eye gaze, body orientation, and so on. (DuBois 1991: 87)

The nature of ViMELF data makes the need for such a development evident. Examples (1) and (2) with Figures 1 and 2 from the ViMELF recordings illustrate the role nonverbal elements can play. In example (1), the German participant SB27 and her Italian conversation partner FL25 talk about the books they own.

(1) Books (03SB27FL25)

SB27:	I have so much books here that I .. bought,
	but .. I can't read them. ((hehe))
FL25:	look, {shifts camera to show bookshelf} {points to bookshelf}

[3] The guidelines are available on the ViMELF homepage at Trier University of Applied Sciences, Germany (ViMELF 2017a).

I mean .. we have dictionaries,

yes, dictionaries,

but ther- but there are also books there somewhere, {makes brushing-away gesture; arm still extended to back}

Figure 1: Pointing gesture in Books (03SB27FL25). Click on image to see the full video sequence

Even just focusing on the visual level and disregarding, for the moment, paralinguistic features, several interesting features can be shown. FL25 shifts her stance by leaning back and out of the screen so the bookshelf in the background is no longer obscured, indicating awareness of her conversation partner's field of view. This shift of orientation is accompanied by the invitation *look* while FL25 moves her laptop computer so that the camera points to the bookshelf, forcing a shift of perspective also for SB27, who responds with a backchanneling smile and nodding, signaling understanding and marking agreement and engagement, all nonverbally. This is then immediately followed by FL25 extending her right hand to point at the bookshelf and the books in it.

In example (2), the German participant SB73 explains Bavarian traditional male dress code but does not recall the word for braces.

(2) Braces (06SB73ST14)

SB73: ... an:d uhm: they have, {lifts head & rolls eyes}
(1.1) how do you call it uhm,
(1.3) uhm .t, [((ehh))] {imitates braces with both hands}
th- it's uhm .t, {imitates braces with both hands}
... uh like a rubber band,
it goes .. [on your trousers],

ST14: [two things],= {imitates braces with both hands}

SB73: =yeah, {imitates braces with both hands}

ST14: right okay, {smiles & nods}

SB73: yeah [I didn’t], {points at herself with both hands}
ST14: [>I don’t know what is-<]
... >I don’t know what it's-< what's the name for it right.
yeah I know what you mean, {closes eyes} ((hehe))
SB73: yeah, {nods}
it hol- holds the trousers .h? {looks down & lifts arms} ((snuffles))

Figure 2: Imitating gesture in Braces (06SB73ST14) with QR video link. Click on image to see the full video sequence

SB73 uses imitative gestures to convey her meaning; the gesture is then mirrored by her Spanish conversation partner ST14. Shared understanding is negotiated and achieved through nonverbal elements without using, at any point, the lexical item that denotes the referent.

4.2. Transcription guidelines for nonverbal elements: Basic guidelines

The complex nature of gesture and other nonverbal elements that we already commented upon raises the question of how to proceed when developing a systematic annotation system. Whether we consider NVE and speech to be overlapping and complementary, or to represent a single system may depend on the type of gesture we analyze. What is true for all interpretations is that NVE represent an essential part of meaning-making in interaction and cannot be ignored when analyzing multimodal data. The approach taken by the ViMELF transcription team focused on four basic guidelines that tie in with the general transcription guidelines discussed above:

(i) the system should take into account the function of the NVE in interaction,
(ii) it should be as systematic as possible and use a regular and predictable syntax that allows quantitative research,
(iii) it should be as descriptive, but also as simple as possible, and
(iv) it should remain adaptable.

In the development of the ViMELF annotation system the compilers refrained, as far as possible, from interpreting NVE during annotation in order to make the transcripts as objective as possible, leaving it to the researchers to draw their own conclusions. At the same time, Hepburn and Bolden's observation on the complex nature of visible behavior is of particular relevance:

> Although the transcription of both talk and visible behavior is necessarily selective, the transcription of visible behavior may be even more so due to the substantial number of parameters. Moreover, visible behavior involving facial expressions, body posture, gestures and gaze can occur in overlap with each other and with talk. (Hepburn and Bolden 2013: 70)

The video component of ViMELF remains available as an integral part of the corpus, so that researchers can return to the raw data in order to supplement the transcript in a context of a more exhaustive multimodal analysis. The annotation system is specifically left open for additions —if salient gestures are observed that are not yet codified, transcribers can easily expand the taxonomy following general guidelines and mark-up syntax.

The ViMELF project team decided to integrate the gestural annotation into the general transcript rather than creating a separate layer for nonverbal elements. In line with keeping the general transcript syntax as simple and readable as possible, nonverbal elements are universally marked with curly brackets, thus: {shrugs}. As mentioned in Section 1, our definition of nonverbal elements includes not only gestures, but also other embodied talk, such as salient head movements and facial expressions, gaze, physical stance shifts, camera shifts, and interaction happening in interlocutors' surroundings, as these can all be considered salient nonverbal contributions to meaning-making.

4.3. Transcription guidelines for nonverbal elements: Development

Similar to the general transcription process, the development of an annotation scheme for nonverbal elements was a data-driven, bottom-up process that integrated continuous feedback by transcribers. Its aim was the classification of salient nonverbal elements in the form of a clear taxonomy. The development process can be divided into four phases:

(i) a survey of existing transcription practices for gestural research,

(ii) a survey of salient nonverbal elements marked by transcribers in pilot transcriptions,

(iii) the formulation of general guidelines for NVE transcription, and

(iv) an inventory of NVE documented in the data as the basis for transcription.

The phases are briefly illustrated below.

(i) Survey of existing transcription practices. The aim of this phase was to establish whether there is a best-practice approach for the annotation of nonverbal elements that can be used or adapted to the multimodal data. While our approach is informed by the more general transcription practices for NVE in CA and interactional sociolinguistics, none of these schemes, as already discussed, fulfils the specified requirements. There is a number of corpora that integrate gestures in their annotation, mostly aligned and integrated into multi-layer display tools. The *Augmented Multi-party Interaction Corpus* (AMI; cf. Carletta *et al.* 2006) and the *SmartKom Multimodal Corpus* (Schiel *et al.* 2002) both use experimental annotation systems that focus on distinguishing conversational (or interactional) and nonconversational gestures with the aim of enhancing machine gesture recognition and are, due to this narrow focus, not suitable for adaptation with ViMELF data. Several corpora use the MUMIN multimodal coding scheme (Allwood *et al*. 2007), which was developed to experiment with annotation of multimodal communication in television data, for example the *Multimodal Human-Computer Interaction Technologies Corpus* (MM HuComTech; cf. Pápay *et al*. 2011). MUMIN focuses on the interpretation of the communicative function of NVE and proposes mutually exclusive categories, "since the focus of the annotation scheme is on the explicit communicative function of the phenomenon under analysis" (Allwood *et. al*. 2007: 278). In other words, "the annotator is asked to select the most noticeable communicative function" (Allwood *et. al*. 2007: 278). This focus on interpretation and the absence of multifunctionality are the key reasons why the system was not considered for use with ViMELF, though the systems share several features, in particular in the basic differentiation of behavior attributes (such as hand shape or head movement).

(ii) Survey of salient NVE marked by transcribers. In a separate pilot phase for the transcription of nonverbal elements, the six project transcribers were asked to transcribe salient NVE in six sample transcripts, constituting about a quarter of ViMELF corpus data, and to describe them in a concise manner. The NVE identified after this round were systematized, resulting in a list of salient NVE and another list of potential inconsistencies regarding descriptive syntax, the concept of salience and the issue of gesture overlap. The transcription team then compared their transcripts and discussed these inconsistencies, contributing to the formulation of general guidelines for transcription.

(iii) Formulation of general guidelines for NVE transcription. In order to make sure transcribers in the main transcription phase would mark gestures consistently, several general principles for transcribing NVE were formulated on the basis of the survey in phase (ii). The general principles formulated during this process are presented in Table 1.

1. Salience	Only salient NVE are transcribed. Salience here refers to NVE contributing to or supporting meaning making, as well as NVE that are referred to on a verbal level or that refer to something that is discussed in the conversation.
2. Markup	All transcribed salient NVE are marked by curly brackets, creating self-contained, searchable markup units that can easily be converted to other data formats such as Extended Markup Language (XML), while being part of an easy to read, lexical transcript.
3. Conciseness and syntax	The transcription syntax of NVE is verb-based and concise. The verb should be in the third person present tense, not the present participle, that is, {nods}, not {nodding}. If there are commonly used verbs that already encompass the NVE they should be used instead of disassembling the NVE into single verb components, such as {makes peace sign} instead of {lifts hand; palm outward} {spreads two fingers}. It is understood that this will always include some level of abstraction and/or interpretation; the aim is not to provide a semiotically precise representation, but an easy to read, concise description.
4. Treatment of consecutive and co-occurring NVE	Consecutive NVE can be transcribed consecutively if no concise transcription exists: {smiles} {nods} {makes thumbs-up gesture}. If several NVE co-occur, they are transcribed in one bracket and connected with &: {smiles & nods}. If the NVE consists of separate stages that could be part of the same NVE, it should not be disassembled into phases but transcribed concisely, e.g. {imitates breathalyzer by blowing into top end of pen}, not {imitates breathalyzer by lifting pen to mouth and blowing into top end of pen and setting it down}.
5. Position and alignment	NVE transcription should follow the intonation unit containing the most salient use, or, if limited to smaller units (e.g. words), follow those units. It is not aligned and not marked for duration, intensity, or speed.
6. Modification	The following modifiers can be added (if co-occurring, in this order): a) direct object, if the main verb does not already contain the object sense (e.g. *nods* includes the object *head* and does not need to be repeated), e.g. *lifts arm*, *lifts hand*. If a NVE is already conventionally named, it may be used as object of *make*, e.g. {makes throwaway gesture} {makes peace sign} {makes air quotes} b) temporal adverb (*three times*, *repeatedly*, ...) c) directional adverb(ial) (e.g. *up*, *down*, *behind ear*), if necessary. Directions are not separately denoted if the specification does not have an influence on the meaning of a NVE —thus, left and right are usually not distinguished. Directions are always given from the speaker's perspective (*outward*, *inward*, *front*, *back* etc.). d) *to* ..., if a target/level needs to be added (*to chair*, *to eye*, etc.) e) *with* ..., indicating for example the hand(s), body parts or objects used in the NVE (e.g. *with left hand*, *with left index finger*), if salient f) *by* [...]-*ing*, if a further modification is needed (e.g. in the case of imitating, as in *imitates breathalyzer by blowing into top end of pen*) g) other additions, for example, if further specification is needed, separated with semicolon {lifts hands; palms outward} {lifts hand; palm up}. This modification should be used sparingly. The sequence should be as short as possible, following the conciseness principle.

Table 1: General principles for the transcription of nonverbal elements in ViMELF (adapted from ViMELF 2017b)

In general, the taxonomy makes use of names of conventionalized Western European NVE as descriptions to reduce complexity in the annotation (e.g. {shrugs}, {nods}). Additional explanations are added in the taxonomy, clarifying NVE that might be culturally specific (e.g. 'peace sign'). The annotation aims to be as descriptive as possible, but in some cases, the terminology used may imply certain meanings or interpret NVE to a certain degree to clarify the context (e.g. 'fist pump'). The actual interpretation of the individual functions of an NVE should remain with the researcher, taking into account the conversational setting, particularly as speakers' different cultural backgrounds increase the probability of diverging functions for similar NVE.

(iv) Inventory of NVE documented in the data. The NVE transcribed during phase (ii) were collected in order to serve as salient examples of NVE that could occur in the transcription guidelines, and to create a data-driven, bottom-up taxonomy for nonverbal elements in ViMELF. The taxonomy was specifically left open so that new salient instances of NVE could be added following the guidelines. This resulted in a taxonomy of salient nonverbal features as presented in Table 2, current as of March 2020.

The taxonomy currently comprises 55 NVE, of which nine are associated with facial expressions, four with the head, including gaze features, three with physical stance and two with the speakers' background. 39 features are associated with hand or body movement; these do not only include movement, but also actions such as standing up, walking, or camera movements that force a shift of perspective. Both guidelines and taxonomy were integrated into the general transcription guidelines and included in the training sessions for student and project transcribers. The current version of guidelines and taxonomy is available online.[4]

[4] Transcription of nonverbal elements (ViMELF 2017b)

Head, including gaze
- Looks (up, down, to side, to upper corner ...)
- Nods (head moves up and down)
- Shakes head (head turns left and right)
- Tilts head (repeatedly)
Facial expressions
- Frowns
- Grimaces (implying negative connotation)
- Purses lips
- Raises eyebrow(s)
- Rolls eyes
- Smiles
- Squints
- Winks
- Yawns
Hands and body
- Claps
- Clasps hands (in front of chest if not otherwise specified)
- Drinks from …
- Drums fingers (rapid movements with fingers)
- Eats …
- Folds arms
- Hits/thumps with … on …
- Holds … to …
- Holds up … (two fingers, glass of wine, etc.)
- Imitates … (drinking, breathalyzer, braces, shape of …, size of …, etc., by ...)
- Lifts hand (to…, or lifts hand; palm up)
- Makes ...
- air quotes (imitates quotation marks with index and middle fingers)
- beat gesture (up and down hand movement during speech; cf. McNeill 1992)
- box gesture (raises hands and moves them, palms vertical, cf. Cassell 1998)
- fist pump (makes fist with one hand, moves fist quickly downwards)
- fist(s)
- okay sign (index finger and thumb together, other fingers extended)
- peace sign (makes a 'V' with index and middle finger, palm outward)
- swiping gesture (moves hand sidewards quickly)
- throwing-away gesture (downward hand movement, palm down, cf. Bressem *et al.* 2013)
- brushing-away gesture (upward hand movement, palm down, cf. Bressem *et al.* 2013)
- thumbs-up gesture, thumbs-down gesture (fist with thumb extended)
- Moves hand to … (to mouth, to forehead, etc.)
- Moves hands … (in circle, outwards, up, etc.)
- Opens hand(s) (outwards movement, palm upwards)
- Points to … (with …) (with index finger/hand, etc.)
- Puts … on …
- Rubs … against ... (rubs thumb against index and middle fingers)
- Scratches (head) (if salient, e.g. in combination with hesitation, thinking, etc.)
- Shifts camera to show …
- Shows … (moves object in front of camera/closer to screen to focus attention)
- Shrugs
- Stands up / sits down

Table 2: Taxonomy of salient nonverbal elements in ViMELF (2017b)

Hands and body (cont.)
- Touches … (head, ear, shoulder, etc.)
- Types
- Walks to ...
- Waves
Physical stance
- Leans … (forward, backward, towards …)
- Sits up (straighter than before)
- Shifts position
Background
- Movement: Noun + verb in third person (roommate walks past screen, etc.)
- Background sounds: Noun + verb in third person (baby cries etc.), noun (clicking sound, etc.) if source is unclear

Table 2 (continuation)

4.4. Transcription guidelines for nonverbal elements: Implementation

Transcription of ViMELF in its current version 1.0 took place over a period of two years, from April 2016 until April 2018. This was due to the fact that recordings of ViMELF data were still ongoing and that transcription phases needed to coincide with research periods of student transcribers. Both transcription guidelines and taxonomy were updated and expanded repeatedly during the transcription period and transcripts checked repeatedly for conformity with the latest version before publication in May 2018.

4.5. Transcription guidelines for nonverbal elements: Advantages and disadvantages

One potential disadvantage of the ViMELF transcription system for nonverbal elements is the selective perspective introduced by the maxims of salience and conciseness.

The salience maxim was a central component in transcriber training in order to ensure a maximum degree of agreement in the transcripts. While this aim was reached with inter-transcriber reliability at more than nine percent at the end of the second transcription phase, it also had the effect that features that were classified as nonsalient were almost uniformly excluded from the transcripts. The resulting transcript is thus necessarily a selective representation of the interactions. Features that are not transcribed but that are of interest to researchers in another context, for example, idiosyncratic or incidental gestures, will need to be extrapolated from the raw audio and video data which is an integral part of the corpus.

To illustrate the consequences of the focus on conciseness, we will consider the transcription of {imitates}, one of the most intriguing subcategories of NVE documented in the corpus. {Imitates} refers to instances where gestures are used to imitate an action, activity, object, shape, size, etc. while explaining or referring to it verbally, as shown in example (2) and Figure 2. There are 130 instances of imitation in ViMELF ranging from imitations of clothing items, actions, states of mind, or physical distances to cultural traditions. Obviously, these can be very complex sequences that are condensed by the transcription team so as to facilitate comparative research. An example is the imitative action by one of the speakers who explains the word *alcohol* by first imitating a breathalyzer and then the action of drinking, as shown in example (3) and Figure 3.

(3) HE19: oh which one? {leans forward}
SB93: alcohol ((/ˈalkɔ:l)). [was that just in Norway],
HE19: [{shakes head once, leans forward}]
what is it, aikai ((/aikai/))? {leans forward}
SB93: so, alcohol ((/ˈalkɔ:l/)),
<alcohol> ((/ˈalkohəʊl/)). {imitates breathalyzer by blowing in top of pen}
[what you drink].
HE19: [oh: the], {scratches head with left hand} the brand?
SB93: nO, what you drink, {imitates drinking}

Figure 3: Imitating gestures in Alcohol (05SB93HE19). Click on image to see the full video sequence

Example (3) shows an imitation sequence in the interaction between a German and a Finnish conversation partner. SB93 asks her Finnish interlocutor about the price of alcohol in Finland, but HE19 does not understand the word *alcohol*, probably because

SB93 pronounces it very quickly and without aspirating the /h/. Since HE19 still does not understand the intended meaning after repeatedly asking for and receiving a repetition of the problematic word, SB93 finally uses gesture to support her meaning making. She blows in the top of a pen as if to imitate a breathalyzer (see Figure 3, first picture, the circle indicates the top of the pen). In this case, the pen at hand is 'recruited' as imaginary breathalyzer, which is then handled accordingly by blowing into the top. When HE19 still does not understand, confusing SB93's pronunciation with the name of the Finnish national alcohol retailer (*Alko*), SB93 adds the combined verbal/nonverbal explanation *what you drink,* {imitates drinking} (see Figure 3, second picture), using her left hand to illustrate the act of drinking. In the subsequent exchange it becomes clear that the negotiation of meaning is successful.

While the conciseness maxim may lead to a simplification of complex sequences, it is also necessary to ensure that the features can be systematically retrieved, which is an important aspect from a corpus analytical perspective.

On balance, we argue that the selective focus and the integrated standardization is what makes the proposed transcription system feasible for use in a corpus linguistic context. A decisive advantage of the proposed system is the quantification of nonverbal elements which allows a mixed-methods approach. Is the effort involved in such a detailed transcription justified in view of its potential for linguistic research? We would argue that despite the considerable time necessary for transcription of NVE, in ViMELF roughly between one and two hours per minute of recorded data, even a comparatively small corpus such as this, with roughly 150,000 tokens, is large enough to quantify selected features (as will be shown in Section 5), and small enough for a meaningful qualitative analysis of multimodal features. The proposed transcription scheme provides considerable benefits: it helps the researcher to find specific instances for closer analysis, and it provides quantitative observations that can be used to guide the analyst's perspective, and that would not be possible to make by close qualitative analysis of the data alone.

The concise (if necessarily less detailed) multimodal transcript and the possibility to access the original data allow a more complete picture of conversational interaction and open up new perspectives on multimodal conversation.

In order to show the research potential of this resource, two approaches are briefly illustrated in Section 5; for a more extensive study of multimodality in ViMELF see Brunner (2021).

5. Using the transcription system

5.1. Quantification of nonverbal elements

One of the main advantages of having a searchable multimodal corpus is the possibility to use quantitative methods to investigate the role of NVE in interaction. A quantitative analysis of NVE in ViMELF (2018) is easy to carry out and provides first insights into how NVE contribute to meaning-making in interaction, and how they correlate with other discourse features. There are 7,449 salient transcribed NVE in ViMELF, distributed over 6,463 instances of transcribed nonverbal behavior (one instance of non-verbal behavior may contain several parallel NVE). Interestingly, only 35 NVE account for 80.4 percent of all transcribed nonverbal behavior, as illustrated in Figure 3. Of those, the six most frequent NVE ({nods}, {shakes head}, {shrugs}, {raises eyebrows}, {tilts head}, and {smiles} already account for 50.9 percent of the total.

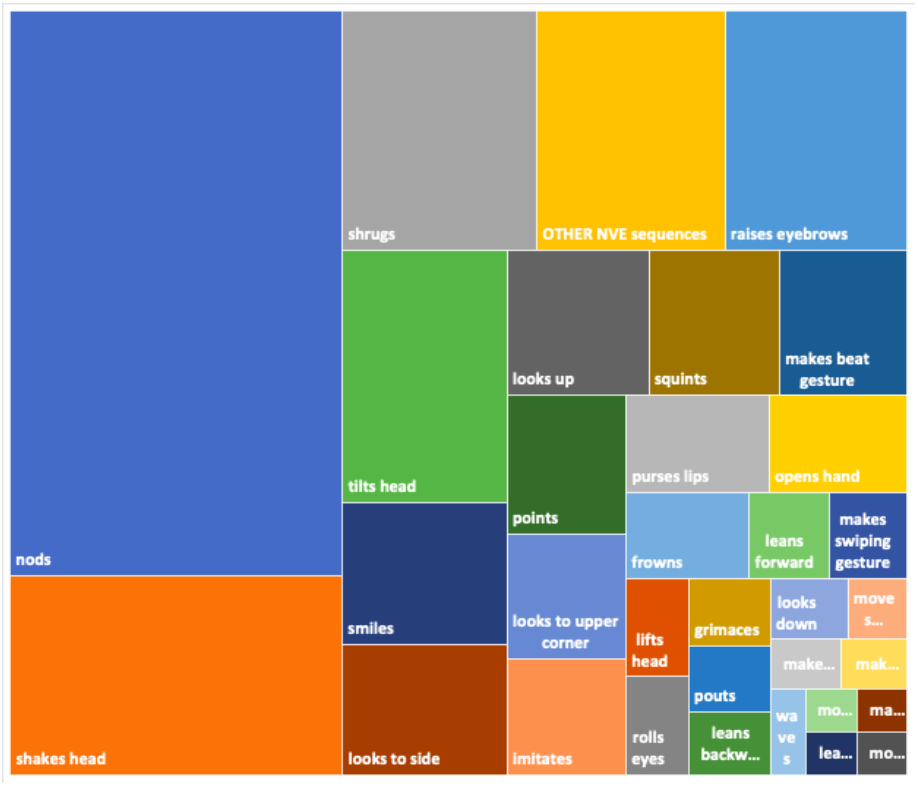

Figure 3: Visualization of relative distribution of nonverbal elements in ViMELF

The descriptive taxonomy advocated in the NVE transcription system means that these instances need to be analyzed qualitatively to determine the function of the respective NVE in the discursive context. If we examine, for example, the 430 instances of {shrugs}, the third frequent NVE, a qualitative analysis reveals a multitude of situational interpretations and functions (see also Brunner *et al*. 2017 for additional examples). {Shrugs} can, for example, express uncertainty (*maybe*. {shrugs}); indicate normalcy and a lack of excitement (*basically the same thing*. {shrugs}); mark a lack of knowledge (*I don't know much about Germany anyway. so.* {shrugs}); mark resignation (*I think it is*, {shrugs} (1.3) *almost impossible*); indicate agreement ({shrugs} *right*); indicate a lack of preferences (*you want to go first or should I?* [...] {shrugs} *go ahead*); signal exasperation (*it doesn't make sense, why is the table female?* {shrugs}); and express disapproval (*and the government* {shrugs} *is not doing anything*).

The example illustrates the complexity of possible interpretations and functions in interaction and shows both the advantages and disadvantages of a mainly descriptive annotation system. On the one hand, it allows the quantification of an additional mode without having to refer to the original data in every case; on the other hand, it will still be necessary to perform a detailed manual analysis of the context. Even in this case, though, relevant instances will be easier to retrieve without going through all of the original recordings.

Another clear advantage lies in the possibility to correlate NVE with other elements and with each other. On a basic level, NVE can correlate with lexical items: the 1,741 instances of {nods}, for example, collocate ($p<0.05$) with *yeah*, *mhm*, *right*, and *okay*, while {shakes head} correlates with *no* and *not*. Both correlations are not surprising. But correlations can become complex very fast: the 386 instances of {tilts head}, for example, correlate with *well*, and *then*, but also with the NVE {nods}, {tilts head}, and {raises eyebrows} as well as the paralinguistic elements ((*ehh*)), ((*heh*)), ((*laughs*)), which are various types of laughter. A correlation analysis like this has the potential to enhance our understanding of meaning making. The gesture {tilts head} clearly is part of a complex negotiation sequence that may include hesitation markers, laughter, and other gestures. It can thus contribute one additional facet to a mixed-method analysis of talk-in-interaction with quantitative and qualitative elements.

5.2. Discursive functions of nonverbal elements

A corpus that is annotated for multimodality also allows researchers to easily extract nonverbal elements and to focus on their broader functions in discourse. One comprehensive recent study uses ViMELF data for the development of a model for multimodal meaning negotiation in video-mediated interactions based on ViMELF data (Brunner 2021). Preliminary results show that although interlocutors are separated by the computer screen and in different environments, they make use of nonverbal elements to complement, replace, nuance, and support their verbal utterances multimodally. Understanding is signaled through both verbal and nonverbal back-channeling. Interlocutors notice aspects of their respective surroundings and can focus attention on them through both verbal means and complementary focusing NVE, for example through pointing, object showings, or camera shifts. Interlocutors also interact with their immediate environment, causing disruptions that have to be negotiated. These first results show the potential for further work with multimodally annotated corpora in order to investigate spoken discourse.

6. CONCLUSION

In our article we propose a concise annotation system for nonverbal elements in spoken discourse and illustrate its application in the context of the ViMELF corpus as a way of integrating unstructured multimodal data into a corpus context. We also show several applications of a corpus annotated with the proposed system for both quantitative and qualitative research on multimodal discourse. The main challenges in creating annotation for multimodal features are (i) the necessity to create systematic criteria for selecting which multimodal features to transcribe and (ii) the need to create an annotation syntax that facilitates systematic quantitative research while preserving a consolidated transcript including lexical, nonverbal, and paralinguistic elements. The resulting taxonomy is based on the two principles of salience and conciseness, and constitutes a systematic, descriptive and comprehensive annotation system. Our aim is not to replace existing approaches, but to provide a robust, easy-to-use tool for the annotation of nonverbal elements as key elements of linguistic meaning-making. In considering both benefits and drawbacks of such a system, we argue that it represents a balanced approach that allows researchers to structure rich, multimodal data and contributes to opening the way for the development of more rich-data corpora and a wide range of applications.

REFERENCES

Adolphs, Svenja and Ronald Carter. 2013. *Spoken Corpus Linguistics: From Monomodal to Multimodal*. London: Routledge.

Allwood, Jens, Loredana Cerrato, Kristina Jokinen, Constanza Navarretta and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41: 273–287.

Bezemer, Jeff and Carey Jewitt. 2010. *Multimodal Analysis: Key Issues.* London: Continuum.

Bressem, Jana, Silva H. Ladewig and Cornelia Müller. 2013. Linguistic Annotation System for Gestures (LASG). In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf eds. *Body-Language-Communication: An International Handbook on Multimodality in Human Interaction.* Berlin: Walter de Gruyter, 1098–1125.

Brunner, Marie-Louise. 2021. *Understanding Intercultural Communication: Negotiating Meaning and Identities in English as a Lingua Franca Skype Conversations*. Saarbrücken: Saarland University PhD dissertation.

Brunner, Marie-Louise, Stefan Diemer and Selina Schmidt. 2017. "... okay so good luck with that ((laughing))?" - Managing rich data in a corpus of Skype conversations. In Turo Hiltunen, Joe McVeigh and Tanja Säily. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* [Studies in Variation, Contacts and Change in English 19]. Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/brunner_diemer_schmidt/ (01 May, 2021.)

Calbris, Geneviève. 2011. *Elements of Meaning in Gesture.* Amsterdam: John Benjamins.

Carletta, Jean, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthatl, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma and Pierre Wellner. 2006. The AMI meeting corpus: A pre-announcement. In Steve Renals and Samy Bengio eds. *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005, Edinburgh, UK, July 11–13, 2005, Revised Selected Papers* (Lecture Notes in Computer Sciences 3869). Berlin: Springer, 28–39. https://link.springer.com/chapter/10.1007/11677482_3

Cassell, Justine. 1998. A framework for gesture generation and interpretation. In Roberto Cipolla and Alex Pentland eds. *Computer Vision in Human-machine Interaction*. Cambridge: Cambridge University Press, 191–215.

Dressler, Richard A. and Roger J. Kreuz. 2000. Transcribing oral discourse: A survey and a model system. *Discourse Processes* 29/1: 25–36.

Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1/1: 71–106.

Edwards, Jane A. 1993. Principles and contrasting systems of discourse transcription. In Jane A. Edwards and Martin D. Lampert eds. *Talking Data: Transcription and Coding in Discourse Research*. Hillsdale: Lawrence Erlbaum Associates, 3–31.

F4transkript. Dr. Dresing & Pehl GmbH. https://www.audiotranskription.de/f4transkript/ (07 May, 2021.)

Goodwin, Charles. 2000. Action and embodiment within situated human interaction. *Journal of Pragmatics* 32/10: 1489–1522.

Goodwin, Charles. 2007. Environmentally coupled gestures. In Charles Goodwin, Susan D. Duncan, Justine Cassell and Elena Levy eds. *Gesture and the Dynamic Dimensions of Language.* Amsterdam: John Benjamins, 195–212.

Goodwin, Marjorie H. and Charles Goodwin. 2000. Emotion within situated activity. In Nancy Budwig, Ina Č. Užgiris and James V. Wertsch eds. *Communication: An Arena of Development*. Stamford: Greenwood Publishing Group, 33–53.

Hepburn, Alexa and Galina Bolden. 2013. The conversation analytic approach to transcription. In Jack Sidnell and Tanya Stivers eds. *The Handbook of Conversation Analysis*. Hoboken: John Wiley & Sons, 57–76.

Jefferson, Gayle. 1973. A case of precision timing in ordinary conversation: Overlapped tag-positioned address terms in closing sequences. *Semiotica* 9/1: 47–96.

Joo, Jungseock, Francis F. Steen and Mark Turner. 2017. Red Hen Lab: Dataset and tools for multimodal human communication research. *KI-Künstliche Intelligenz* 31/4: 357–361.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Richie Key ed. *The Relationship of Verbal and Nonverbal Communication*. Berlin: Mouton de Gruyter, 207–228.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance.* Cambridge: Cambridge University Press.

Kerswill, Paul and Ann William. 2002. "Salience" as an explanatory factor in language change: Evidence from dialect levelling in urban England. In Mari C. Jones and Edith Esch eds. *Language Change: The Interplay of Internal, External and Extra-Linguistic Factors.* Berlin: Mouton de Gruyter, 81–110.

Kress, Gunther. 2011. Multimodal discourse analysis. In John P. Gee and Michael Handford eds. *The Routledge handbook of discourse analysis*. London: Routledge, 35–50.

McNeill, David. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, David. 2008. *Gesture and Thought.* Chicago: University of Chicago press.

McNeill, David. 2017. *Brief Introduction to Annotation.* http://mcneilllab.uchicago.edu/analyzing-gesture/intro_to_annotation.html (01 May, 2021.)

McNeill, David and Susan Duncan. 2000. Growth points in thinking-for-speaking. In David McNeill ed. *Language and Gesture.* Cambridge: Cambridge University Press, 141–161.

Mondada, Lorenza. 2014. Pointing, talk, and the bodies. In Mandana Seyfeddinipur and Marianne Gullberg eds. *From Gesture in Conversation to Visible Action as Utterance: Essays in Honor of Adam Kendon*. Amsterdam: John Benjamins, 95–124.

Norris, Sigrid. 2002. The implication of visual research for discourse analysis: Transcription beyond language. *Visual Communication* 1/1: 97–121.

Pápay, Kinga, Szilvia Szeghalmy and István Szekrényes. 2011. Hucomtech multimodal corpus annotation. *Argumentum* 7: 330–347.

Sacks, Harvey, Emanuel A. Schegloff and Gail Jefferson. 1978. A simplest systematics for the organization of turn taking for conversation. In Jim Schenkein ed. *Studies in the Organization of Conversational Interaction*. New York: Academic Press, 7–55.

Schiel, Florian, Silke Steininger and Ulrich Türk. 2002. *The SmartKom Multimodal Corpus at BAS.* München: Ludwig-Maximilians Universität München Press.

Scollon, Ron and Philip LeVine. 2004. Multimodal discourse analysis as the confluence of discourse and technology. In Philip LeVine and Ron Scollon (eds.), *Discourse*

and Technology: Multimodal Discourse Analysis. Washington: Georgetown University Press, 1–6.
Streeck, Jürgen. 2009. *Gesturecraft: The Manu-facture of Meaning.* Amsterdam: John Benjamins.
ViMELF. 2017a. *ViMELF Transcription Conventions.* Birkenfeld: Trier University of Applied Sciences. http://umwelt-campus.de/case-conventions (01 May, 2021.)
ViMELF. 2017b. *Transcription of Nonverbal Elements.* Birkenfeld: Trier University of Applied Sciences. https://www.umwelt-campus.de/fileadmin/Umwelt-Campus/SK-Weiterbildung/Dateien/Transcription_of_non-verbal_elements_in_CASE.pdf (01 May, 2021.)
ViMELF. 2018. *Corpus of Video-Mediated English as a Lingua Franca Conversations.* Birkenfeld: Trier University of Applied Sciences. http://umwelt-campus.de/case (01 May, 2021.)

Corresponding author
Marie-Louise Brunner
Trier University of Applied Sciences
Environmental Campus Birkenfeld
P.O. Box 13 80
55761 Birkenfeld
Germany
e-mail: ml.brunner@umwelt-campus.de

received: February 2020
accepted: May 2021

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

RiCL Research in Corpus Linguistics

The burden of legacy: Producing the *Tagged Corpus of Early English Correspondence Extension* (TCEECE)

Lassi Saario[a] – Tanja Säily[a] – Samuli Kaislaniemi[b] – Terttu Nevalainen[a]
University of Helsinki[a] / Finland
University of Eastern Finland[b] / Finland

Abstract – This paper discusses the process of part-of-speech tagging the *Corpus of Early English Correspondence Extension* (CEECE), as well as the end result. The process involved normalisation of historical spelling variation, conversion from a legacy format into TEI-XML, and finally, tokenisation and tagging by the CLAWS software. At each stage, we had to face and work around problems such as whether to retain original spelling variants in corpus markup, how to implement overlapping hierarchies in XML, and how to calculate the accuracy of tagging in a way that acknowledges errors in tokenisation. The final tagged corpus is estimated to have an accuracy of 94.5 per cent (in the C7 tagset), which is circa two percentage points (pp) lower than that of present-day corpora but respectable for Late Modern English. The most accurate tag groups include pronouns and numerals, whereas adjectives and adverbs are among the least accurate. Normalisation increased the overall accuracy of tagging by circa 3.7pp. The combination of POS tagging and social metadata will make the corpus attractive to linguists interested in the interplay between language-internal and -external factors affecting variation and change.

Keywords – corpus annotation; corpus markup; spelling normalisation; TEI-XML; part-of-speech tagging; Late Modern English

1. INTRODUCTION[1]

1.1. Legacy corpora

Many of the corpora used to study the history of English have a history of their own. That is especially true of the pioneering corpora from the early 1990s that are still used nowadays, such as the *Helsinki Corpus of English Texts* (HC) and *A Representative*

[1] People who worked in the tagging project besides the authors deserve to be mentioned here. Mikko Hakala, Emanuela Costea, Anne Kingma and Anna-Lina Wallraff were responsible for a large part of the semi-manual normalisation. We would also like to thank Paul Rayson, Jukka Suomela, Turo Hiltunen, Arja Nurmi and Gerold Schneider for their assistance and advice. This work was supported in part by the Academy of Finland, Grants 293009 and 323390.

Corpus of Historical English Registers (ARCHER). They were originally encoded in a format that was state-of-the-art at the time, but as the years have gone by, the original format has become outdated and incompatible with new tools. This is a common problem among old corpora and the reason why they are called 'legacy corpora'. While most of them are small in size by present-day standards, the vast amount of qualitative work invested in them still makes them valuable compared to today's big data corpora which put quantity before quality (see Hundt and Leech 2012; Davies 2019). Legacy corpora deserve to be rescued, then, but how?

The solution is, of course, to convert them into a new format (as has been done to the HC and ARCHER that have been converted into TEI-XML), but that solution is bound to cause new problems. The original compilers of legacy corpora cannot have foreseen the needs of their successors, and the choices made by them in the past (such as the markup schemes chosen or the features omitted from the texts) limit the options available in the present. If the corpus has been based on secondary sources such as printed editions of original manuscripts, the interpretative work done by the editors also sets certain preconditions. In a sense, it might seem easier to start the markup process from scratch. If these problems can be solved, however, the conversion can bring an old corpus back to life again. Not only may the new format be richer than the old one, but it may also allow for further enrichment (e.g. new kinds of annotation) and so broaden the scope of possible research questions that the corpus can shed light on, making it even more valuable than it was before.

In this article, we present a case study of one legacy corpus and the problems related to converting and enriching it, some of which are general while others are specific to legacy corpora. Our case in point is part-of-speech tagging the *Corpus of Early English Correspondence Extension* (CEECE). The CEECE could be classified as a second-generation legacy corpus, as it follows the markup conventions of the HC but comes equipped with substantially richer metadata. It will serve as an example of a legacy corpus that has been successfully 'rescued' and enriched with annotation.

1.2. Enrichment of the CEECE

The CEECE opens a window into the sociohistorical variation and change of Late Modern English (LModE) through personal letters, sampled and digitised from published editions

(see Kaislaniemi 2018). Plenty of successful studies have been conducted on the corpus since the initiation of its compilation in 2000 (see e.g. Nevalainen *et al.* 2018). Until now, however, the letter texts have remained in largely unstructured form. More sophisticated queries require more structured data where the linguistic features of interest, such as parts of speech, are explicitly annotated. We hope the need for richer data will now be satisfied as we present the new POS tagged version of the corpus, known as the *Tagged Corpus of Early English Correspondence* (TCEECE).

In the original CEECE, text files are accompanied by an external database that contains structured metadata about the letters and the correspondents, whereas the actual letter bodies consist of mostly unstructured text. While the tagging project did not increase the amount of data (defined as the word count), it did enrich the data by both structuring the unstructured and adding more structure on top of the pre-existing. First, the texts were converted into TEI-XML so as to make their internal structure more transparent and well-formed. Second, the texts were tokenised into word elements and, third, each token was assigned a POS tag. From this point of view, POS tagging the CEECE illustrates how a small corpus can be made more valuable —perhaps even more valuable than a bigger corpus which is not as rich.

On the other hand, the enrichment caused complications that had to do with, for example, normalising spelling variation, converting the legacy format and calculating the accuracy of the tagging. We believe these to be common problems among corpus annotators, especially those who are working with historical material or trying to update legacy corpora to the 'third generation' (see Hiltunen *et al.* 2017: §3). We hope our experiences will be of use to colleagues wrestling with similar difficulties. We would like the production of the TCEECE to set an example, not only of how heavy the burden of legacy can be but also of how that burden can eventually be overcome.

We will begin with an overview of the history of the corpus, the POS tagging project and the technologies behind it (Section 2). We will then outline the workflow of the project and reflect on critical points (Section 3), followed by a discussion where we look at our choices in retrospective, trying to learn from our mistakes and to come up with suggestions on better policies for others to follow (Section 4). We will conclude with a summary of what we have done and what remains to be done (Section 5).

2. BACKGROUND

2.1. The CEEC family of corpora

The *Corpora of Early English Correspondence* constitute a digitised corpus family compiled by the Helsinki-based *Sociolinguistics and Language History* team to facilitate systematic sociolinguistic research into the history of the English language. It has grown over the years from the original core corpus of 2.6 million words to a family of subcorpora twice that size.

The original version, the *Corpus of Early English Correspondence* (CEEC), was completed in 1998 and covers the period from circa 1410 to 1681. A half-a-million-word *Sampler* version of the corpus (CEECS) was published in 1999, and the corpus at large in 2006. Due to copyright restrictions, this grammatically annotated published version, the *Parsed Corpus of Early English Correspondence* (PCEEC), is slightly smaller than the original one, comprising 2.2 million words. The original version was supplemented by circa 400,000 words of additional material from 1402 to 1663, packaged as the CEEC *Supplement* (CEECSU, unpublished). Later, the corpus team also extended the CEEC into the eighteenth century, creating the CEEC *Extension* (CEECE), a 2.2-million-word subcorpus, which stretches the timeline covered to 1800, earning the corpus family the acronym CEEC-400 as it covers four centuries (see Table 1).

	CEEC	CEECS	PCEEC	CEECE	CEECSU	CEEC-400[2]
Words	2,597,957	450,082	2,159,132	2,218,520	441,304	5,221,349
Collections	96	23	84	77	19	191
Letters	6,053	1,124	4,970	4,923	857	11,714
Writers	778	194	666	308	95	1,125
Time span	c. 1410–1681	1418–1680	1410–1681	1653–1800	1402–1663	1402–1800

Table 1: The CEEC corpus family

2.2. Choice of tagger

The system of grammatical annotation of the CEEC has a history of its own, which is longer and more complex than that of the CEEC corpus family itself. The PCEEC

[2] CEECS and PCEEC are not counted in the numbers of CEEC-400, being subsets of CEEC. Of the two versions of the Plumpton collection, the newer one (in CEECSU) has been excluded from the total counts.

annotation was carried out by the CEEC team in collaboration with researchers from the University of York, with Arja Nurmi in Helsinki being responsible for the part-of-speech tagging and Ann Taylor at York for the syntactic parsing. To ensure compatibility of diachronic corpora that cover largely the same time period, the same annotation system was chosen as had been used earlier to tag and parse the grammatically annotated versions of the HC, that is, the *Penn-Helsinki Parsed Corpus of Middle English* (PPCME2; Kroch *et al*. 2000) and the *Penn-Helsinki Parsed Corpus of Early Modern English* (PPCEME; Kroch *et al*. 2004), which both followed the guidelines of the *Penn Treebank*.[3]

The question of annotation system arose again when plans were made to provide the CEECE with POS tagging. One relevant alternative was to adopt the *Brill* tagger and the *Penn Treebank* tagset used in the *Penn Parsed Corpora of Historical English* and, by doing so, to provide continuity with the POS tagging of the PCEEC. The other alternative, originally also experimented with the HC (Kytö 1996: 5), was to opt for the *Constituent-Likelihood Automatic Word-Tagging System* (CLAWS). This had become the *de facto* standard for corpora made available through the widely used Lancaster University *CQPweb* interface (Hardie 2012), including many Present-day English (PDE) corpora as well as the *Early English Books Online* corpus and the *Corpus of English Dialogues*.[4]

The choice between the two systems depended on a number of factors. As LModE is in many ways close to PDE, comparability between the TCEECE and CLAWS-tagged PDE corpora such as the *British National Corpus* (BNC) and the Brown family of corpora[5] was thought to be advantageous; other LModE corpora had been tagged using various annotation systems, so there was no one model to follow there (Hundt 2014: 2). In terms of tagger performance, the accuracy of the *Brill* tagger on the PCEEC was circa 80–90 per cent (Arja Nurmi, personal communication), which is similar to that of CLAWS on Early Modern English (EModE), although automatic spelling normalisation as a pre-processing step has been shown to improve the CLAWS output (Rayson *et al.* 2007; Hiltunen and Tyrkkö 2013). When applied to present-day corpora, both annotation systems are reported to reach comparable levels of accuracy (c. 96–97%).[6]

Our final decision was reached by considering one more factor, namely the annotation scheme. The *Penn* tagset employed in the PCEEC, designed to be used

[3] See https://www.ling.upenn.edu/hist-corpora/, https://catalog.ldc.upenn.edu/docs/LDC95T7/cl93.html
[4] https://cqpweb.lancs.ac.uk/
[5] https://varieng.helsinki.fi/CoRD/corpora/BROWN/
[6] See http://ucrel.lancs.ac.uk/claws/. For a comparison of the two tagsets, see Lu (2014: 42–47).

throughout the long diachrony of English, has significant drawbacks compared to CLAWS for the study of more modern forms of English. Analysing noun ratios in the PCEEC, Säily *et al.* (2011) found, for example, that the adverb *likewise* was tagged conservatively as a combination of an adjective and a noun (ADJ+N), identically to the noun *gentleman*. Moreover, the annotation scheme follows Huddleston and Pullum's (2002) analysis of prepositions, collating subordinators and prepositions into a single category, which precludes studying them separately unless the corpus is syntactically parsed (Säily *et al.* 2017: 46). As no syntactic parsing was being planned for the CEECE and, unlike in the PCEEC project, checking all the annotation manually was not an option, CLAWS was chosen as the basis for producing the TCEECE.

2.3. Other technological choices

Once CLAWS had been chosen as the tagger, we had yet to choose from the various tagsets that were available for CLAWS. The prominent options at the time were C5 (62 tags), C7 (137–152 tags) and C8 (170 tags).[7] C7 was an enriched version of C5, and C8 likewise of C7. The native output of CLAWS followed C7 and could automatically be mapped to C5, while enrichment into C8 would have required post-processing by a separate software, *Template Tagger* (Fligelstone *et al.* 1997). For that reason, as well as the fact that the BNC had been tagged using C5 and the BNC sampler using C7, we ended up choosing between C5 and C7.

We found it an advantage of C7 that there was a distinct tag for almost every personal pronoun, while C5 only had one tag for all of them (cf. Säily *et al.* 2017: 46). On the other hand, C7 had unnecessarily fine-grained noun categorisation. We decided to provide the tagged corpus in both tagsets, as the C5 tagging could be derived from the C7 tagging without any cost. Neither did we need to check the accuracy of the two taggings separately, for the checking of C7 could also be directly translated into that of C5. Since the BNC Sampler had been tagged in C7, we could largely rely on the same guidelines in checking the accuracy (see Section 4.1 for a comparison of accuracy between the tagsets).

The original markup of the CEECE (as the CEEC-400 in general) is based on that of the HC (Kytö 1996: §3.3.2; Nurmi 1998: §2), which ultimately dates back to the COCOA program that was used on punched cards and magnetic tapes in the 1960s and

[7] See http://ucrel.lancs.ac.uk/claws/

1970s (see e.g. Russell 1965; Corcoran 1974). Before the corpus could be tagged by CLAWS, it had to be converted into XML. The HC had already been converted into TEI P5 XML (Marttila 2011), so it was only natural that we converted the CEECE into a similar schema (see Section 3.2). The BNC, too, had been converted into TEI-XML and made available on *CQPweb*, which encouraged us to import the TCEECE into *CQPweb* as well.

3. Workflow: Problems and solutions

A thorough documentation of the TCEECE project has been published in the *Corpus Resource Database* (Saario and Säily 2020). Figure 1 illustrates the production process. Instead of redocumenting the process in every detail, we will here focus on the central problems we faced, the solutions we came up with and the lessons we learned from them. Many critical choices had to be made, some of which turned out to have a significant effect on the later working stages and the use of the final corpus. Those choices and their effects, as well as the alternative paths that might (and maybe should) have been taken, will be discussed in more depth in Section 4.

Figure 1: A visualisation of the workflow

3.1. Normalisation

The historical spelling variation in the CEECE was normalised to better comply with present-day standards, so as to make it easier for CLAWS to tag. The first stage of normalisation was performed semi-automatically with UCREL's *Variant Detector* (VARD; Baron 2011a, 2011b) as a part of an earlier project, the creation of the *Standardised-spelling Corpora of Early English Correspondence* (SCEEC). The tagged output of VARD includes both original and normalised-spelling variants inside XML-like tags. The normalised form appears in between the XML tags, while the original variant is kept inside an `orig` attribute:

(1) `I would desire you to send me an Oxford <normalised orig="almanack" auto="true">almanac</normalised>`

However, it was the untagged output that was moved on to the next stage of normalisation, so that only the normalised forms remained. While omitting the tags did make the text easier to process, losing the original spellings actually 'impoverished' the data rather than enriched it, conflicting with the ideal expressed in the introduction. Our choice to produce a POS-tagged version of the corpus that was silently normalised (but retained text-level markup elsewhere) was a compromise between maximal annotation and ease of use. The latter consideration applied both to the people involved in the production process —many of whom were research assistants with no knowledge of XML— and to the end-users of the corpus. Leaving out the original spelling was also not seen as a major issue because the CEEC family of corpora was never designed for the study of orthographic variation. The compilers used original-spelling editions to ensure that the linguistic content would be reliable for morphosyntactic studies, but even these editions frequently normalise features such as *u*/*v* variation, capitalisation or punctuation. Recent work has shown that the CEEC compilers' reservations towards using the corpora to study spelling, capitalisation, punctuation or word division were largely warranted (Sairio *et al.* 2018; but see Kaislaniemi *et al.* 2017). In any case, the original (editorial) spelling is preserved in the original version of the corpus, so with access to both versions, users are still able to check the spelling, albeit with some difficulty (see Section 4.2 below).

Further normalisation was performed partly manually and partly automatically. Given the variability of historical spelling, even after being processed with VARD, the CEECE texts contained great numbers of tokens not found in PDE. As we did not have the resources to manually normalise all remaining non-standard items, it was decided to focus on the most frequent types, and ones that were easy to identify. The bulk of these were abbreviations, which are commonly marked by punctuation (`Ld.` for 'Lord'; `desir'd` for 'desired'), superscripts (coded in CEECE with equal signs: `w=ch=` for *w*^{ch} 'which'), or special characters (changed in CEECE to tildes: `com~and` for 'command'; `lr~es` for 'letters'; `p~mit` for 'permit'). Some of the abbreviations in CEECE are still current in PDE, such as *Mrs*, but with formatting that makes them opaque to CLAWS, such as `M=rs=`. In the case of abbreviations not found in PDE, `Sep=br=`, `Sep=t=`, `Septem` and `7=br=` may be intelligible to human readers, but not to CLAWS. And the same applies to otiose abbreviations, mostly marked with superscripts, such as `you=r=` 'your' and the ubiquitous `y=e=` 'the' —which was particularly tricky when occurring without superscripts, as it needed to be disambiguated from the plural pronoun *ye*. This

variability in the spelling and formatting of abbreviations in the CEECE partly reflects manuscript reality, but also the practices of different editors and printers.

In the first cycle of post-VARD normalisation, a concordancer was used to find such items. These were then manually reviewed in a spreadsheet, and those chosen for normalisation were given normalised forms in a separate column. Finally, Python scripts were used to replace the original variants in the texts with the normalised forms. Nearly 8,000 abbreviated words or otherwise non-standard variants were normalised in this way. In the second cycle, the same process was repeated by a different method: a sample of the twice-normalised texts from across the CEECE was run through CLAWS, and problematic items were identified. Scripts were then used to capture and normalise such cases in the whole corpus, to a number of roughly 9,200. Aside from abbreviations, other frequent features requiring such manual attention included punctuation as well as word division in indefinite pronouns (*every body* > *everybody*) and reflexive pronouns (*my self* > *myself*) (see Saario and Säily 2020: §3). The total number of (semi-)manual replacements came to 17,024.

More information about the original text was, of course, lost at this stage, as the variants to be normalised were simply replaced with PDE forms without leaving any trace of the original variants. All text-level encoding that was involved in the original variant was also lost in the process, so that, for example, `fin[{is{]h'd`, where `[{is{]` marks an emendation, was normalised into `finished` where there is no sign of the original spelling nor the emendation. Again, getting rid of that information did streamline the pipeline but it also had unfortunate consequences for the use of the end product, which will be discussed in Section 4.2.

3.2. XML conversion

Throughout the normalisation process, the corpus remained in the ancient COCOA format. The parameter lines that preceded each letter were not a problem, but the letter bodies involved a great deal of custom text-level coding that CLAWS would not have understood (see Saario and Säily 2020: §4.4). Apart from paragraph shifts that were only implicitly indicated, there were ‘P-lines’ to mark page shifts and various code brackets to mark comments, emendations, etc.[8] The easiest solution would have been to remove all

[8] Special characters (e.g. the pound sign) had already been converted into XML in the normalisation.

text-level coding, which would have lost still more information and further impoverished the data. We wanted to avoid that outcome and decided to convert all the coding into XML in order for it to survive through POS tagging.

Our approach to XML could be characterised as ‘modest’ in the sense of Hardie (2014). While we did model our XML schema after that of the HC (Marttila 2011) which, in turn, is based on the TEI guidelines, we did not even try to implement all of their potential but only the bare minimum that was required to preserve the encoded information. We also prioritised effectiveness over tidiness and sought to automatise the conversion as far as possible. Despite the modesty of our intentions, several problems arose along the way, the most symptomatic two of which are treated here.[9]

3.2.1. Separating ‘proper comments’ from ‘emendation comments’

Following the HC, editors’ comments in the CEECE were originally annotated with the code `[\...\]` and compilers’ comments with `[^...^]`. One issue was that both codes were used for two different types of annotations. The same code might be used in, for example, the following two instances:

(2) `reminding him of his obligations and his [\ONE WORD MISSING\]`

(3) `she walked about [\her\] Chamber`

The difference is that in the first instance the comment is a meta-level remark about the body text, whereas in the second instance it is an editorial addition that is meant to be read as a part of the text like an emendation (which are encoded as `[{...{]`). We call the two uses a ‘proper comment’ and an ‘emendation comment’, respectively.

The two uses of the same code had to be recognised and separated in order for CLAWS to ignore proper comments and only tag emendation comments, which are in effect normalisations. The task was performed by an algorithm, based on the observation that proper comments, unlike emendation comments, generally involved several

[9] Soon after completing the first version of the TCEECE, we got funding for converting the entire CEEC-400 into XML (see Saario 2020). This allowed us to further develop our converter program and update the underlying XML format of the TCEECE accordingly. We here describe the updated format.

consecutive capital letters. The latter were placed between XML tags, while the former were hidden inside XML attributes, as follows:

(4) `<note resp="editor" value="ONE WORD MISSING" />`

(5) `<note resp="editor">her</note>`

We acknowledge that our algorithm is not perfect, as it assumes the original encoders to have been more consistent in their application of the codes than they probably were, but it does succeed frequently enough to justify itself. It is more important to extract relevant structure than to avoid casual errors. Hardly any such errors have shown up yet, and they can be manually corrected whenever they do.

3.2.2. Dealing with ‘trans-token’ codes

In addition to editors’ comments, compilers’ comments and emendations, there were separate codes for headings, typeface changes and foreign language, encoded as `[}...}]`, `(^...^)` and `(\...\)`, respectively. Whenever a code covered a single token or a sequence of tokens, it could be converted directly into XML, as in the examples above. Problems arose when a code transcended the token division, as in (6).

(6) `thank you for the unus[{ual plea{]sure it has given me.`

The obvious XML translation would have been `unus<supplied>ual plea</supplied>sure`, but CLAWS only tags whole words (cf. the nesting problem in Section 3.3.1). If the information about the exact range of the code was to be kept, it had to be done indirectly. We initially decided to extend the corresponding XML code into the closest sequence of whole words and keep the original encoded sequence inside an ‘orig’ attribute (cf. the treatment of ‘split’ words in Rodríguez-Puente *et al.* 2019: 73), as shown in (7).

(7) `<supplied orig="unus[{ual plea{]sure">unusual pleasure</supplied>`

Later, following a suggestion by our colleagues in Lancaster, we added a ‘range’ attribute to record the start and end indices of the code. The characters in each sequence were indexed starting from zero (skipping whitespaces). This approach also generalised into cases where there are several code ranges in one sequence, as in (8).

(8) `<note resp="editor" range="1,4;5,7" orig="m[\ist\]r[\es\]s">mistress</note>`

If matters were not complicated enough, sometimes the consecutive codes were of different kinds —and not only could there be several consecutive codes in one sequence, but there could also be codes inside codes, and more codes inside those codes. In the end, we did find a way to contain all this variation and convert it systematically into XML, but it required a robust algorithm and an elaborate conceptualisation of the hierarchy of codes (Saario 2021).

3.3. Tokenisation and POS tagging

The XML edition of the CEECE was tokenised and POS tagged by CLAWS, using the C7 tagset, and post-processed by a simple script that switched the POS tags inside foreign language passages (encoded as `<foreign>...</foreign>` in XML) to the proper tag for foreign words. The final output was then converted into various formats in both C7 and C5. The accuracy of C7 tagging was checked from a sample and mapped to that of C5.

3.3.1. Final format

The direct output of CLAWS is called ‘vertical’ as there is one line for each token. Long tokens and XML tags with whitespaces have been moved to an associated supplement file and must be manually retrieved from there when the output is converted back to XML. The conversion was performed by a separate program written by Paul Rayson.[10]

We would have liked to enclose sentence tokens in `s` elements and word tokens in `w` elements, as in, for example, the BNC XML edition. Unfortunately, the text-level codes that had been translated into XML before tagging turned out to be incompatible with `s` elements. Whenever the converter reached an opening XML tag inside a sentence, it closed the `s` element before the tag even if the sentence continued after it, for example:

[10] https://github.com/UCREL/convert

(9)
```
<s>
   <w id="410.1" pos="PPH1">It</w>
   <w id="410.2" pos="VM">will</w>
   <w id="410.3" pos="VBI">be</w>
</s>
<supplied range="4,9" orig="some[{thing{]">
   <w id="410.4" pos="PN1">something</w>
</supplied>
<w id="410.5" pos="JJ">chargeable</w>
```

In the above example, there is no reason why the `s` element could not continue over the `supplied` element, but there are other cases where a supplied passage (or another text-level code) continues across a sentence break. In those cases, closing the former sentence and opening the latter one at the breaking point would not have been well-formed XML, as the supplied passage would have been nested in neither sentence. One solution would have been to split the `supplied` code at the sentence break, which would have required some robust post-editing. We took the easiest path and simply omitted `s` elements. The sentence tokenisation is, nevertheless, implicitly encoded in the `id` attributes of `w` elements, where the first number identifies the sentence and the second identifies the word in that sentence. We acknowledge that our solution is not optimal and hope that a better one will be found in the future (cf. the problem of overlapping hierarchies in Marttila 2014: 200–201).

3.3.2. Checking the accuracy

The value of a tagged corpus largely depends on the accuracy of tagging. Typically, the accuracy is estimated by calculating certain key figures from a representative sample. The overall accuracy rate is the number of correctly tagged tokens divided by the total number of tokens. Each tag also has two measures: ‘precision’ is the number of correct assignments divided by the total number of assignments of the tag, and ‘recall’ is the number of correct assignments divided by the total number of tokens for which the tag in question is correct (see Figure 2 for illustration). All it takes to calculate those figures is to go through the tokens one by one and determine for each whether the tag assigned by the tagger is correct, and if it is not, what the correct alternative is.

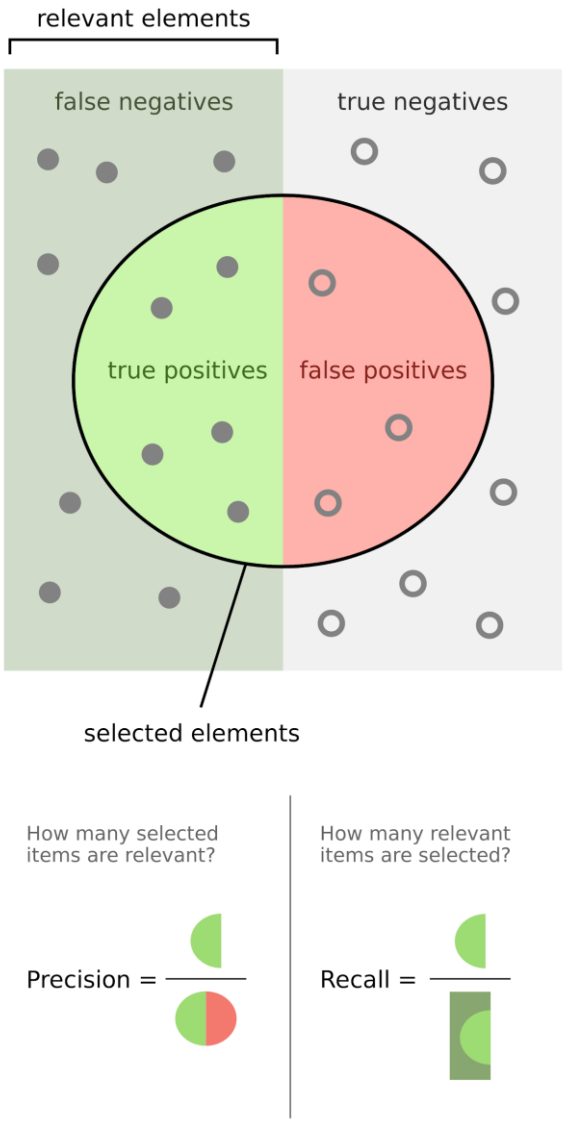

Figure 2: Precision and recall[11]

What was said above presupposes that the underlying tokenisation itself is correct, which in the case of the TCEECE was not true. While our XML converter did merge lines of text into paragraphs, tokenisation into sentences and words was left to the tagger. Even if we would have liked to check the tokenisation before tagging, we were unable to do that, as the tagger only produced one output where the text had been both tokenised and tagged. As a result, there are many incorrect tags due to incorrect tokenisation, as in (10).

(10) `and_CC when_RRQ twill_NN1 be_VBI better_JJR`

[11] Source: https://commons.wikimedia.org/wiki/File:Precisionrecall.svg, published by 'Walber' under the licence CC BY-SA 4.0.

Here, we have an obsolete contraction *'twill* that remains non-normalised due to a lack of resources and because it is not marked by any of the features listed in Section 3.1. The tagger has interpreted it as one token while, in fact, there are two tokens (*it* and *will*) that ought to be tagged separately. Similarly, there are cases where one token has been mistaken for two. In these cases, it is senseless to ask what the correct tags for the incorrectly tagged tokens would be, as there are no correctly tagged tokens in the first place.

We solved the problem as follows. Whenever there is one token that should have been two, it is counted as one incorrectly tagged token. Whenever there are two tokens that should have been one, they are counted as two incorrectly tagged tokens (unless the other is a punctuation mark, in which case only that one counts as incorrect). The correct alternatives for the incorrect tags in either case are classified as 'excluded'.

This workaround allowed us to calculate the overall accuracy as well as precisions and recalls for particular tags in a way that does not distort the figures too much. We could, of course, have chosen the opposite way and counted the true tokens (*it* and *will*) instead of those given by the tagger (*twill*), which might have been closer to the presupposition of perfect tokenisation and the idea of having a baseline or 'gold standard' of tagging, against which the actual tagging is measured (Rayson *et al.* 2007: 8). A third alternative would have been to exclude the incorrect tokens from our sample, but we preferred our figures to reflect the errors in tokenisation as well as in tagging. The results of our calculations are presented in the next section.

4. Discussion

In this section, we will reflect on our choices and their effects, trying to assess the use value of the end product and come up with alternative or additional actions that might have improved it. We will first look at the accuracy of tagging from various points of view and compare it with, for example, other tagged corpora. We will then discuss on a more general level the management of corpus projects and outline suggestions based on the lessons we learned.

4.1. Accuracy of tagging

To check the accuracy of tagging, we compiled a sample of 15 letters, comprising 5,245 running words (c. 0.24% of the total word count) that had been tagged using the C7 tagset. The sample is representative in the sense that the average length of letters and the distributions of letter-writers' genders and ranks as well as times of writing somewhat correspond to those in the entire corpus. Post-processed passages of foreign language were excluded from the sample to avoid bias (see Saario and Säily 2020: §6.1). The tagging of the sample was checked, and the accuracy of tagging calculated following the procedure explained in Section 3.3.2. The results are provided in what follows (see also *ibid*.: §§6.4–6.6).

4.1.1. Overall accuracy

The sample had been tokenised by CLAWS into 5,889 tokens, 5,566 of which we classified as accurately tagged. The overall accuracy of the sample is therefore 94.5 per cent. There is a great deal of variation among the letters, however: the lowest accuracy is 90.6 per cent while the highest is 97.2 per cent. Of the 323 inaccurately tagged tokens in the sample, 32 (9.9%) were due to incorrect tokenisation, which allows us to conclude that the accuracy of tokenisation is 99.5 per cent.

Having combined the results with metadata on the letters and their writers, we learned that the accuracy is 95.4 per cent for letters by men and 92.8 per cent for letters by women, which might be explained by the fact that women typically had less access to education than men. Neither is it unexpected that the accuracy is 93.5 per cent for letters from the seventeenth century and 94.7 per cent for letters from the eighteenth century, given that spelling in English became increasingly standardised over that time. There is no observable difference in the overall accuracy between the upper and lower social ranks, but that is understandable as the sample only has three letters from the lowest rank. In general, the lower social ranks are underrepresented in our corpus for obvious reasons and those who are represented are often the most literate ones, which is bound to cause some bias.

The letter with the lowest tagging accuracy (90.6%) was written by Joanna Clift, a domestic servant with no formal education. We have reason to believe that the Clift letter collection is, in fact, one of the worst collections in terms of tagging accuracy, as it

contains relatively many letters from poorly literate writers (see Saario and Säily 2020: §6.6). We had no time to manually correct its tagging because of its size, but we did correct the smaller Pauper collection which we also expected to have been tagged rather inaccurately for the same reason (*ibid*.). We learned that the accuracy of the uncorrected Pauper collection was 87.9 per cent, which may be considered the approximate lower bound for the tagging accuracies of all CEECE collections.

4.1.2. Accuracy by tags

In addition to the overall accuracies, end-users of the tagged corpus will want to know the accuracies of particular tags, especially those they intend to use in their research. Precision and recall were calculated for each C7 tag (see Saario and Säily 2020: Appendix 2) and are summed up into groups in Table 2.

Tag group		Selected assignments (a)	Relevant assignments (b)	True assignments (c)	Precision (c / a)	Recall (c / b)
	Punctuation marks	570	562	562	98.6%	100.0%
A-	Articles	443	439	438	98.9%	99.8%
C-	Conjunctions	384	393	359	93.5%	91.3%
D-	Determiners	179	168	158	88.3%	94.0%
I-	Prepositions	535	527	511	95.5%	97.0%
J-	Adjectives	259	265	237	91.5%	89.4%
M-	Numbers	103	96	95	92.2%	99.0%
N-	Nouns	1,032	1,017	944	91.5%	92.8%
P-	Pronouns	639	644	636	99.5%	98.8%
R-	Adverbs	385	412	358	93.0%	86.9%
TO	Infinitive marker	106	108	106	100.0%	98.1%
V-	Verbs	1,140	1,123	1,063	93.2%	94.7%
	Miscellaneous[12]	114	103	99	86.8%	96.1%

Table 2: The precisions and recalls of C7 tags grouped into categories[13]

[12] Includes, for example, negation, genitive marker, letters of the alphabet and existential *there*.

[13] Note that for each group of tags, the values (a)–(c) are sums of those of the tags in the group. The value (c) might be greater were the particular tags mapped onto the level of the groups, which would consequently improve precision and recall. For example, in the group of nouns, there are 73 false negatives, 46 of which

The most accurate tag groups in terms of both precision and recall are articles, punctuation marks, pronouns and infinitive markers. Numbers also have a high recall even if their precision is relatively low. Determiners and miscellaneous tags are worst in precision, adjectives and adverbs in recall.

What this means in practice is that queries for precise tags print concordances where most lines truly represent the tags in question, but users cannot trust that most true instances are included unless the recall is high, too. On the other hand, concordances for tags with a high recall but low precision include many false instances but, at least, users may suppose most true instances are included and they just have to eliminate the false ones, which is often easier than to dig up missing instances from outside the search results. We might go as far as to say that in corpus linguistics, recall is generally more important than precision (cf. Hoffmann 2005: 21).

To help users to deal with tags that have low recall, we have also calculated the accuracies by pairs of true and false tags (Saario and Säily 2020: Appendix 2). If one were interested in, for example, the tag JJ (general adjective), one would not only know that the recall is 88.9 per cent, and thus 11.1 per cent of all true JJs in the sample have been tagged as something else; one would also know that 3.8 per cent of them have been tagged as VVN, 2.6 per cent as NN1 and so on, which would help to trace the missing JJs.

The specification by tag pairs reveals that some tags have often been confused with other tags under the same group: this is the case, for instance, when a general adverb (RR) has been mistaken for a degree adverb (RG). That is, of course, more forgivable an error than misplacement in a completely wrong category. If the tagset would not distinguish between the two kinds of adverbs, the given case would not cause an error. It is therefore instructive to see what happens to precision and recall when the C7 tagging is mapped into the more coarse-grained C5.

Surprisingly enough, the mapping does not increase the overall accuracy by more than 0.2 percentage points (pp). Of all the 323 errors, only 11 go away. While a transition into C5 significantly impoverishes the annotation, it barely improves the accuracy; specifically, it does not suffice to overcome the poor recall of adjectives and adverbs, even if the latter is slightly increased (by 1.0pp). That does not prevent the tagging from

---

are confusions between different noun tags. The number of true group assignments (as opposed to particular tag assignments) is therefore 944 + 46 = 990.

being useful, however, as long as the users recognise it does not represent the "God's truth" (Rissanen 1989: 17).

4.1.3. Accuracy by version

Next, we shall compare the accuracy of tagging across the stages of normalisation to find out how much the tagging was improved by each stage. Let us call the original CEECE 'Version 0', the VARD-processed (or 'VARDed') corpus 'Version 1' and the further normalised corpus 'Version 2'. Above, we have already discussed the accuracy of Version 2, having converted it into XML and tagged by CLAWS. The accuracies of the other two versions were determined in the same way: a sample was compiled from the earlier versions of the same letters, converted and tagged, and the tagging was then checked following the same principles as with the final corpus. The results are summarised in Figure 3, where maximum and minimum are the accuracies of the most and least accurately tagged letters, respectively.

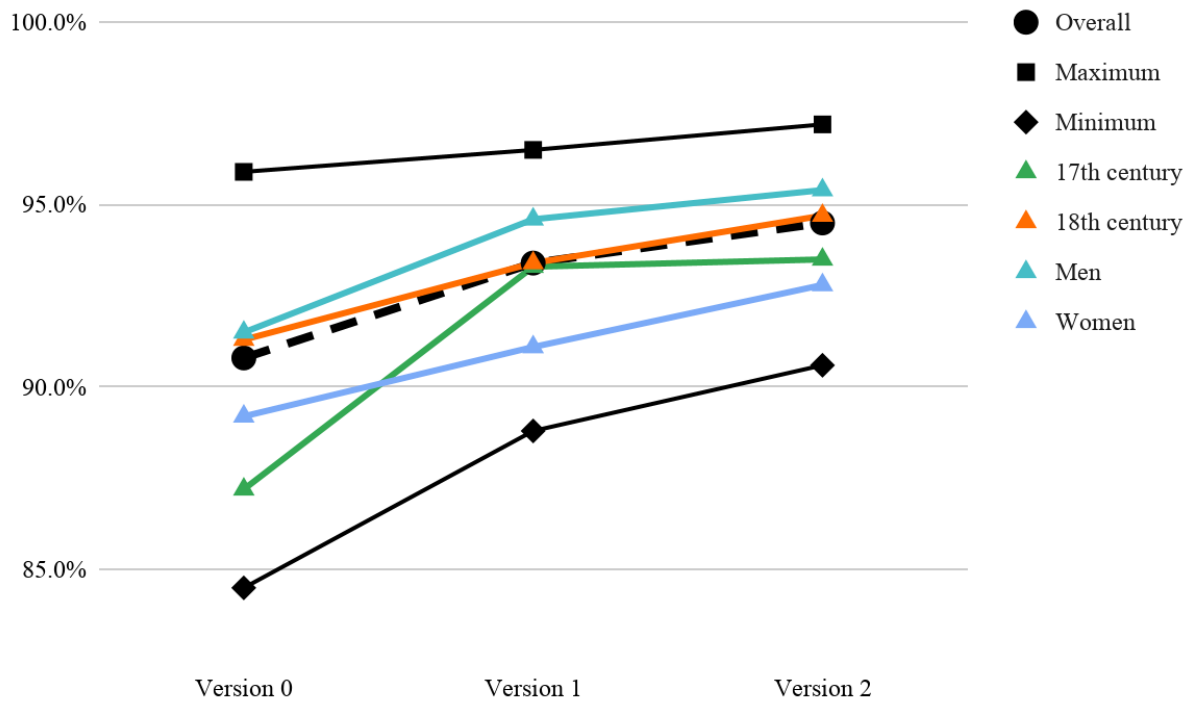

Figure 3: The tagging accuracy of the corpus versions 0, 1 and 2

The overall accuracy of the original corpus (Version 0) was 90.8 per cent. Normalisation in VARD increased the accuracy by 2.6pp to 93.4 per cent and further normalisation by

1.1pp to 94.5 per cent. The difference between maximum and minimum accuracy (the 'range') decreased from 11.4 per cent to 7.7 per cent and finally to 6.6 per cent.

By contrast, Rayson *et al.* (2007) tested the effect of normalisation on CLAWS tagging with two samples from EModE: one from Shakespeare's plays and one from the *Lampeter Corpus of Early Modern English Tracts*.[14] They observed that Shakespeare's initial accuracy of 81.94 per cent was increased to 84.81 per cent (+2.9pp) in VARD and to 88.88 per cent (+4.1pp) in full manual normalisation. For the *Lampeter* sample, which dates from a later period (1640s) and is stylistically closer to the kind of data CLAWS is familiar with, the figures were 88.46 per cent, 89.39 per cent (+0.9pp) and 93.22 per cent[15] (+3.8pp), respectively. The accuracies are lower than those of the TCEECE because of the earlier language form, but the changes in accuracy are more comparable. Differences in the effect of VARDing might have something to do with genre, as the speech-related genres of plays and correspondence probably involve more spelling variation than tracts and pamphlets. Differences in further normalisation are interesting, as the TCEECE was not fully normalised like Shakespeare and *Lampeter*; still, the 1.1pp improvement in the former is relatively good compared to the circa 4pp improvement in the latter.

VARDing the CEECE increased the seventeenth-century accuracy by 6.1pp and the 18th-century accuracy by 2.1pp, bringing the former to almost the same level as the latter. That is not surprising, given that VARD has been designed for EModE in particular. Further normalisation did not improve the seventeenth century by more than 0.2pp, but it did improve the eighteenth century by 1.3pp and so compensated for the bias of the earlier stage. Yet, the sample only has two letters from the seventeenth century, so one must be careful not to generalise too much.

Another (albeit slighter) difference between the two stages of normalisation concerns the gender of writers. The accuracy of men's letters was increased in VARD by 3.1pp and women's by 1.9pp. Further normalisation, in turn, increased men's accuracy by 0.8pp and women's accuracy by 1.7pp. This might imply that men's letters are easier to normalise (semi-)automatically, based on general patterns of variation, whereas women's letters require closer attention to the idiosyncrasies of individual writers.

[14] http://korpus.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM

[15] In the calculation of this figure, a passage of Latin which CLAWS had failed to tag as FWs (foreign words) was excluded from the sample, just like we did with our sample. The figure without exclusion is 91.24 per cent.

4.1.4. Comparison with ARCHER

Schneider *et al.* (2016) employed CLAWS to tag a sample of ARCHER that had been normalised by VARD. They originally used the C5 tagset and mapped it to the *Penn* tagset which only has 39 tags. The accuracy of the final tagging is reported to be 87.8 per cent in the seventeenth century and 93.2 per cent in the eighteenth century, which is 5.8pp and 1.7pp lower than the respective accuracies of the TCEECE in C5. Were the accuracies for the TCEECE calculated in the *Penn* tagset, the difference with respect to ARCHER would be even larger.

The difference between ARCHER and the TCEECE in the seventeenth century is expected, as the TCEECE only covers the end of the century. One must also bear in mind that ARCHER was not further normalised beyond VARDing like the TCEECE. A closer comparison requires that we determine the accuracy of the eighteenth-century part of the TCEECE as it was after VARDing and before further normalisation, using the C5 tagset, and compare it to the eighteenth-century part of ARCHER. We get the result that the TCEECE accuracy is 93.7 per cent, that is, 0.5pp higher than ARCHER. This is surprising, since private spelling as represented by the TCEECE is typically more variable than the spelling of published texts, which is what ARCHER mostly represents. Perhaps, the ARCHER genres have presented the tagger with challenges of a different sort, such as mathematical formulae.

Comparison with other corpora presupposes that the accuracy figures have been calculated in the same way. We have tried to be as transparent as possible about our principles of calculation (see Section 3.3.2), but earlier research has been somewhat vague on the matter. In addition to the treatment of tokenisation errors, one ambiguity that should be resolved is the role of punctuation marks. In checking the tagging of ARCHER, punctuation marks were counted as tagged tokens (Gerold Schneider, personal communication). While we have also followed this convention for comparability, we wonder if it really is wise to equalise punctuation marks with other tokens, given that their tagging tends to be correct by default and is not very interesting anyway.

Even if the tagging of the TCEECE is relatively accurate, it is useful to know what more could have been done to improve it. In the corpus manual, we have listed plenty of known issues, some of which could have been prevented by additional normalisation whereas others are more difficult to avoid before tagging (Saario and Säily 2020: §7).

Comprehensive post-processing by UCREL's *Template Tagger* and manual correction of more collections are, of course, options that we may consider in the future.

4.2. Ideal of gradual enrichment

As we have already noted, the ideal of enrichment did not actualise throughout the process. In the spelling normalisation, information was lost on the original variants as well as text-level coding in normalised variants. Secondly, as already noted, the corpus was not tokenised until it was tagged by CLAWS, so the token identifiers of the TCEECE cannot be used to refer back to earlier versions of the CEECE. Thirdly, because of the problems noted, the TCEECE is largely unsynchronised with the non-tagged, non-tokenised, non-converted or non-normalised versions of the CEECE which will still be used and developed alongside the tagged corpus. For instance, when users find an interesting passage in the TCEECE and want to check its original spelling, they are unable to directly identify the same tokens in the original CEECE. They do have the letter identifier that helps them to find the original letter, but from there onwards, they are on their own, trying to discern the corresponding tokens from the unstructured text that may look a lot different than its normalised, reformatted, tokenised and tagged counterpart.

On the other hand, some people may consider it a relief that not all layers of annotation are piled on top of each other in one file. If they were, users of the corpus might feel overloaded with information, struggling to discern the relevant parts from the thick jungle of code. It is for this reason that, for example, the compilers of CHELAR decided to keep the POS tagged and TEI-XML versions of their corpus separate (Rodríguez-Puente *et al.* 2019: 79–80). Indeed, it seems as if there were an upper bound on how far the enrichment of a corpus should go. Corpus developers should be careful about enriching one corpus version too much. If you add too much annotation, the corpus will become unusable and lose its value.

As much as we sympathise with the underlying concern, we believe there is something to be done other than just settling for many imperfect versions. The layers of annotation can be separated to distinct files by means of stand-off markup that preserves their linking to the 'primary text' (see Marttila 2014: 195ff). On the other hand, even if it is not optimal for an end-user to have all the data in one place, that does not mean there could not be such a place where the plainer versions come from. A distinction should be

made between an all-inclusive ‘master’ version, maintained by developers of the corpus, and simplified subalternate versions that are actually used by researchers. If the master corpus is encoded in XML, tailor-made versions can be derived from it with XSL transformations to suit each user’s individual needs (see e.g. the BNC stylesheets).[16]

There is an additional reason for separating maintenance and use. Our experiences with the TCEECE and other CEEC corpora have taught us that the maintenance of many parallel versions of one corpus becomes excessively laborious as time goes by. The more versions there are, the likelier it is that changes are made to some versions while others fall out of sync (for a cautionary example, see Saario 2020: §1). It would be preferable to have all the data kept up to date in one branch and have a version control system (e.g. *Git*)[17] keep track of the changes.

Of course, that still leaves open the question of how exactly the master corpus is to be constructed, organised and encoded. Incorporating multiple overlapping layers of annotation into one format is a challenge that will not be solved here (see Marttila 2014: §5.6). In an ideal world, we would have been able to envision an all-inclusive format right from the start, allowing us to do things in a more logical order. First, we would have converted the original CEECE into XML. Second, we would have tokenised the unstructured text, which would have assigned identifiers to the data points that could then have been referred back to from later stages. Third, we would have normalised the spelling, keeping the original tokens in store alongside the normalised ones. Finally, we would have POS tagged the normalised tokens. Each stage would have built on the earlier ones, and no information would have been lost in the process. See Figure 4 for illustration (and cf. Figure 1).

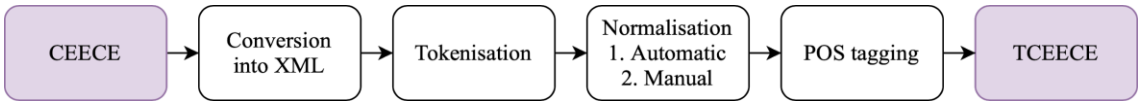

Figure 4: A visualisation of the ideal workflow

That said, we acknowledge our vision will be inaccessible to many corpus compilers for similar reasons as it was to us. Linguists often do not have the expertise to do things like implementing layered annotation, writing XSL transformations or using a version control system, and nor do many of the temporary research assistants who do a major part of the

[16] http://www.natcorp.ox.ac.uk/using/index.xml?ID=stylesheets
[17] https://git-scm.com

actual work. Even if they did, the third-party software they use (e.g. VARD or CLAWS) might not support the ideal workflow without adjustments. In the real world, people will have to come up with easy workarounds past difficult problems, just like we did. Still, it is useful to evaluate different workarounds against the ideal way of doing things, as it may help to avoid the worst pitfalls.

5. Conclusion

The long legacy of the CEECE is still present in the TCEECE. The markup has changed, but the content of the letters is still based on the source editions from which they have been compiled. Traces of the legacy format remain in, for example, attributes and headers. Yet the new format is well-formed and valid XML that largely complies with the TEI guidelines and is compatible with modern tools, which we hope is enough to prolong the life of the corpus by decades.

The value of the TCEECE may be measured along various axes. Extensive automatisation throughout the production process has resulted in errors that should be manually corrected. The accuracy of tagging seems sufficient, even if it could be improved by more ambitious post-processing. The end product is not as rich as it could be, which some users may find a good thing, while the maintainers will have to face the fact that we now have one more parallel version of the same corpus. Yet the corpus is primarily intended to provide a resource that can be easily used by linguists, including those with little technical know-how. All in all, what we have accomplished so far may very well be a good enough compromise between the desiderata of effectiveness, correctness, richness, usability and maintenance.

At the time of writing, the TCEECE is being imported to our *CQPweb* server and will hopefully soon be available to researchers and visitors in our unit. Preliminary research on neologisms has already been tried on the corpus; other prospective topics include large-scale investigations of variation and change in POS frequencies (cf. Säily *et al.* 2011, 2017) as well as keyness and collocation analyses that take word class into account. The combination of POS tagging and social metadata in a relatively large and representative historical corpus of private writing will make the corpus attractive to many linguists interested in the interplay between language-internal and -external factors affecting language variation and change (excluding orthography). In the future, we may

consider further enriching the corpus by adding, for example, lemmatisation to word tokens or distinct markup to the formulaic elements of letters.

References

ARCHER = *A Representative Corpus of Historical English Registers*. 1990–1993/2002/2007/2010/2013. Originally compiled under the supervision of Douglas Biber and Edward Finegan at Northern Arizona University and University of Southern California; modified and expanded by subsequent members of a consortium of universities. https://www.projects.alc.manchester.ac.uk/archer/ (25 February, 2020.)

Baron, Alistair. 2011a. VARD 2. Computer program. http://ucrel.lancs.ac.uk/vard/ (25 February, 2020.)

Baron, Alistair. 2011b. *Dealing with Spelling Variation in Early Modern English Texts*. Lancaster: Lancaster University dissertation. https://eprints.lancs.ac.uk/id/eprint/84887/ (25 February, 2020.)

BNC = *The British National Corpus*, version 3 (BNC XML edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. http://www.natcorp.ox.ac.uk (25 February, 2020.)

CEEC-400 = *Corpora of Early English Correspondence*. 2020. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, 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/ (19 June, 2021.)

CEECE = *Corpus of Early English Correspondence Extension*. 2012. 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/ (19 June, 2021.)

CLAWS. Computer program. Developed by UCREL at Lancaster University. http://ucrel.lancs.ac.uk/claws/ (25 February, 2020.)

Corcoran, Paul E. 1974. COCOA: A FORTRAN program for concordance and word-count processing of natural language texts. *Behavior Research Methods & Instrumentation* 6/6: 566.

Davies, Mark. 2019. Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In Carla Suhr, Terttu Nevalainen and Irma Taavitsainen eds. *From Data to Evidence in English Language Research* (Language and Computers 83). Leiden: Brill, 66–87.

Fligelstone, Steve, Mike Pacey and Paul Rayson. 1997. How to generalize the task of annotation. In Roger Garside, Geoffrey Leech and Anthony McEnery eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman, 122–136. http://ucrel.lancs.ac.uk/papers/CAB_CH08.pdf (25 February, 2020.)

Hardie, Andrew. 2012. *CQPweb* – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17/3: 380–409.

Hardie, Andrew. 2014. Modest XML for corpora: Not a standard, but a suggestion. *ICAME Journal* 38: 73–103.

HC = *The Helsinki Corpus of English Texts*. 1991. Compiled by Matti Rissanen (Project leader), Merja Kytö (Project secretary); Leena Kahlas-Tarkka, Matti Kilpiö (Old

English); Saara Nevanlinna, Irma Taavitsainen (Middle English); Terttu Nevalainen, Helena Raumolin-Brunberg (Early Modern English). Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/ (19 June, 2021.)

Hiltunen, Turo, Joe McVeigh and Tanja Säily. 2017. How to turn linguistic data into evidence? In Turo Hiltunen, Joe McVeigh and Tanja Säily eds. *Big and Rich Data in English Corpus Linguistics: Methods and Explorations* (Studies in Variation, Contacts and Change in English 19). Helsinki: VARIENG. https://varieng.helsinki.fi/series/volumes/19/introduction.html (19 June, 2021.)

Hiltunen, Turo and Jukka Tyrkkö. 2013. Tagging Early Modern English Medical Texts (1500–1700). Presentation at *The First Corpus Analysis with Noise in the Signal Workshop* (CANS 2013), 22 July, Lancaster University, UK. http://ucrel.lancs.ac.uk/cans2013/abstracts/Hiltunen%20Tyrkk%C3%B6.pdf (25 February, 2020.)

Hoffmann, Sebastian. 2005. *Grammaticalization and English Complex Prepositions: A Corpus-based Study*. London: Routledge.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Hundt, Marianne ed. 2014. *Late Modern English Syntax*. Cambridge: Cambridge University Press.

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. *The Oxford Handbook of the History of English*. Oxford: Oxford University Press, 175–188.

Kaislaniemi, Samuli. 2018. The *Corpus of Early English Correspondence Extension* (CEECE). In Terttu Nevalainen *et al*. eds., 45–59.

Kaislaniemi, Samuli, Mel Evans, Teo Juvonen and Anni Sairio. 2017. 'A graphic system which leads its own linguistic life'? Epistolary spelling in English, 1400–1800. In Tanja Säily *et al*. eds., 187–214.

Kroch, Anthony, Ann Taylor and Beatrice Santorini. 2000. *The Penn-Helsinki Parsed Corpus of Middle English.* Department of Linguistics: University of Pennsylvania.

Kroch, Anthony, Beatrice Santorini and Lauren Delfs. 2004. *The Penn-Helsinki Parsed Corpus of Early Modern English.* Department of Linguistics: University of Pennsylvania.

Kytö, Merja. 1996. *Manual to the Diachronic Part of The Helsinki Corpus of English Texts: Coding Conventions and Lists of Source Texts* (third edition). Helsinki: Department of English, University of Helsinki. http://clu.uni.no/icame/manuals/HC/INDEX.HTM (25 February, 2020.)

Lu, Xiaofei. 2014. *Computational Methods for Corpus Annotation and Analysis*. New York: Springer.

Marttila, Ville. 2011. *Helsinki Corpus TEI XML Edition Documentation*. Helsinki: VARIENG. https://helsinkicorpus.arts.gla.ac.uk/display.py?fs=100&what=manual (25 February, 2020.)

Marttila, Ville. 2014. *Creating Digital Editions for Corpus Linguistics: The Case of Potage Dyvers, a Family of Six Middle English Recipe Collections*. Helsinki: University of Helsinki dissertation. http://urn.fi/URN:ISBN:978-951-51-0060-3 (25 February, 2020.)

Nevalainen, Terttu, Minna Palander-Collin and Tanja Säily eds. 2018. *Patterns of Change in 18th-century English: A Sociolinguistic Approach*. Amsterdam: John Benjamins.

Nurmi, Arja ed. 1998. *Manual for the Corpus of Early English Correspondence Sampler, CEECS*. Helsinki: Department of English, University of Helsinki. http://korpus.uib.no/icame/manuals/CEECS/ (25 February, 2020.)

PCEEC = *Parsed Corpus of Early English Correspondence*. 2006. Annotated by Arja Nurmi, Ann Taylor, Anthony Warner, Susan Pintzuk, and Terttu Nevalainen. Compiled by the CEEC Project Team. York: University of York and Helsinki: University of Helsinki. http://hdl.handle.net/20.500.12024/2510 (25 February, 2020.)

Rayson, Paul, Dawn Archer, Alistair Baron, Jonathan Culpeper and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Matthew Davies, Paul Rayson, Susan Hunston and Pernilla Danielsson eds. *Proceedings of Corpus Linguistics 2007, 27–30 July, University of Birmingham, UK*, article 192. http://ucrel.lancs.ac.uk/publications/CL2007/ (25 February, 2020.)

Rissanen, Matti. 1989. Three problems connected with the use of diachronic corpora. *ICAME Journal* 13: 16–19.

Rodríguez-Puente, Paula, Cristina Blanco-García and Iván Tamaredo. 2019. Mark-up and annotation in the *Corpus of Historical English Law Reports* (CHELAR): Potential for historical genre analysis. *Journal of the Spanish Association of Anglo-American Studies* 41/2: 63–84.

Russell, D. B. 1965. COCOA ―A Word-Count and Concordance Generator. http://www.chilton-computing.org.uk/acl/applications/cocoa/p001.htm (25 February, 2020.)

Saario, Lassi. 2020. *Conversion of the CEEC-400 into XML. A Manual to Accompany the XML Edition*. Helsinki: VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/xml_doc.html (19 June, 2021.)

Saario, Lassi. 2021. *XmlConverter. A Java Application to Process the File Format of the Corpora of Early English Correspondence*. Helsinki: VARIENG. https://version.helsinki.fi/ceec/ceec-tools/XmlConverter (19 June, 2021.)

Saario, Lassi and Tanja Säily. 2020. *POS Tagging the CEECE. A Manual to Accompany the Tagged Corpus of Early English Correspondence (TCEECE)*. Helsinki: VARIENG. https://varieng.helsinki.fi/CoRD/corpora/CEEC/tceece_doc.html (19 June, 2021.)

Säily, Tanja, Terttu Nevalainen and Harri Siirtola. 2011. Variation in noun and pronoun frequencies in a sociohistorical corpus of English. *Literary and Linguistic Computing* 26/2: 167–188.

Säily, Tanja, Turo Vartiainen and Harri Siirtola. 2017. Exploring part-of-speech frequencies in a sociohistorical corpus of English. In Tanja Säily *et al*. eds., 23–52.

Säily, Tanja, Arja Nurmi, Minna Palander-Collin and Anita Auer eds. 2017. *Exploring Future Paths for Historical Sociolinguistics*. Amsterdam: John Benjamins.

Sairio, Anni, Samuli Kaislaniemi, Anna Merikallio and Terttu Nevalainen. 2018. Charting orthographical reliability in a corpus of English historical letters. *ICAME Journal* 42/1: 79–96.

SCEEC = *Standardised-spelling Corpora of Early English Correspondence*. 2012. Compiled by Terttu Nevalainen, Helena Raumolin-Brunberg, Samuli Kaislaniemi, Jukka Keränen, Mikko Laitinen, Minna Nevala, Arja Nurmi, Minna Palander-Collin, Tanja Säily and Anni Sairio. Standardised by Mikko Hakala, Minna Palander-Collin and Minna Nevala. Department of English / Department of Modern Languages, University of Helsinki. https://varieng.helsinki.fi/CoRD/corpora/CEEC/ (19 June, 2021.)

Schneider, Gerold, Marianne Hundt and Rahel Oppliger. 2016. Part-of-speech in historical corpora: Tagger evaluation and ensemble systems on ARCHER. In Stefanie Dipper, Friedrich Neubarth and Heike Zinsmeister eds. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)* (Bochumer Linguistische Arbeitsberichte 16). Bochum: Ruhr-Universität Bochum, 256–264. https://www.linguistics.rub.de/konvens16/proceedings.html (25 February, 2020.)

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/ (19 June, 2021.)

TEI Consortium, eds. 2020. *Guidelines for Electronic Text Encoding and Interchange*. Last updated on 13 February, 2020. http://www.tei-c.org/P5/ (25 February, 2020.)

Corresponding author
Lassi Saario
P.O. Box 24
FI-00014
University of Helsinki
Finland
e-mail: lassi.saario@helsinki.fi

received: February 2020
accepted: June 2021

RiCL Research in Corpus Linguistics

How to prepare the video component of the *Diachronic Corpus of Political Speeches* for multimodal analysis

Camille Debras
Université Paris Nanterre / France

Abstract – The *Diachronic Corpus of Political Speeches* (DCPS) is a collection of 1,500 full-length political speeches in English. It includes speeches delivered in countries where English is an official language (the US, Britain, Canada, Ireland) by English-speaking politicians in various settings from 1800 up to the present time. Enriched with semi-automatic morphosyntactic annotations and with discourse-pragmatic manual annotations, the DCPS is designed to achieve maximum representativeness and balance for political English speeches from major national English varieties in time, preserve detailed metadata, and enable corpus-based studies of syntactic, semantic and discourse-pragmatic variation and change on political corpora. For speeches given from 1950 onwards, video-recordings of the original delivery are often retrievable online. This opens up avenues of research in multimodal linguistics, in which studies on the integration of speech and gesture in the construction of meaning can include analyses of recurrent gestures and of multimodal constructions. This article discusses the issues at stake in preparing the video-recorded component of the DCPS for linguistic multimodal analysis, namely the exploitability of recordings, the segmentation and alignment of transcriptions, the annotation of gesture forms and functions in the software ELAN and the quantity of available gesture data.

Keywords – DCPS; multimodal political discourse analysis; gesture studies

1. INTRODUCTION

Still under construction, the *Diachronic Corpus of Political Speeches* (henceforth DCPS) is a collection of 1,500 full-length political speeches in English. It includes speeches delivered worldwide by English-speaking male and female elected politicians in various settings (such as election speeches, parliamentary or party conference speeches, inaugural addresses) from 1800 up to the present time. Speech transcripts are being enriched with semi-automatic morphosyntactic annotations, in the form of lemmatisation and part-of-speech tagging with *TreeTagger* (Schmid 1994). Transcripts are also supplemented with discourse-pragmatic manual annotations including audience responses (Heritage and Greatbatch 1986), speech openings and closures. The DCPS is

designed to fulfil the following criteria: achieve maximum representativeness and balance for political English speeches from major national English varieties in time, preserve detailed metadata, as well as enable corpus-based studies of syntactic, semantic and discourse-pragmatic variation and change on political corpora.

For speeches given from 1950 onwards, video-recordings of the original delivery are often retrievable online. This opens up avenues of research in multimodal linguistics, in which studies on the integration of speech and gesture in the construction of meaning (Kendon 2004; Norris 2004; Müller *et al*. 2013, *inter alia*) can include analyses of recurrent gestures (Müller *et al.* 2013) and of multimodal constructions (Steen and Turner 2013; Zima and Bergs 2017).

This article discusses the practical and methodological issues at stake in preparing the video-recorded component of the DCPS for linguistic multimodal analysis, including the analysis of gesture. It focuses on four main areas:

1) Exploitability of the recordings.
2) Preparing the spoken component of the data: transcription, segmentation of the transcription and alignment of the transcription with the video.
3) Preparing the visual component of the data: annotation of gesture forms and their functions.
4) Quantity of data: how much is "enough"?

After a short presentation of current issues in multimodal discourse analysis, I develop the issues at stake in these four areas in separate sections, before moving on to a general conclusion on the perspectives for innovative diachronic multimodal analyses of political discourse based offered by the DCPS.

2. MULTIMODAL (POLITICAL) DISCOURSE ANALYSIS

2.1. A multimodal approach to the study of discourse: A focus on the interaction between speech and gesture

Spoken communication is multimodal by nature (Norris 2004, *inter alia*). Meaning-making relies on the integration of multiple modes of communication which belong to two different modalities of communication: the oral-aural and the kinesic-visual modalities. The multimodal approach to (political) discourse analysis proposed here focuses on the interplay between actions in speech and gesture that are coordinated in

time (Kendon 2000; Mondada 2016). The contribution of speech can be further subdivided into the verbal mode (discourse, at the segmental level) and the vocal mode (at the suprasegmental level: prosodic phenomena including intonation, volume and speed of the delivery), alongside the gestural mode (Ferré 2011, 2019). Visible bodily actions, including gestures, are so closely intertwined with speech in the construction of meaning that they can be considered part of language themselves (Kendon 2000; Müller *et al.* 2013, *inter alia*). Alongside multimodality, a second main feature of spoken communication is sequentiality: actions done with speech or the body are inscribed in time, one after the other or simultaneously, and they take on their meanings and functions as part of this simultaneous and sequential unfolding of actions and mobilisation of resources. A gesture can thus be defined as a bodily action that "(belongs) to the 'story line' of the interaction" (Kendon 1986: 6), and that is inscribed in the sequentiality of the interaction, namely that coincides with other actions in the construction of meaning, rather than being there by mere coincidence (Schegloff 1984).

The study of gesture is fundamentally interdisciplinary (Stam and Ishino 2011). Since the analysis of a gesture lends itself to a large range of approaches in various domains (e.g. psychology or anthropology), a recent body of work has developed a specifically linguistic approach to gesture (cf. Müller *et al.* 2013), showing how gestures, traditionally relegated to the para-verbal, actually do lend themselves to linguistic analysis. Although not prototypically linguistic, gestures can be linguistic to some extent (Cienki 2017), notably in terms of forms, functions, and form-function pairings, for instance in the case of recurrent gestures (Ladewig 2014).

The multimodal study of political discourse is a thriving field, which includes the contribution of gestures to discourse structuring and framing (cf. Streeck 2008; Cienki 2009; Wehling 2009; Cienki and Giansante 2014; Debras and L'Hôte 2015, *inter alia*). The video component of the DCPS will provide scholars with new opportunities to supplement this field with studies on (diachronic) variation.

2.2. ELAN, a tool for the study of multimodal data

Several software tools can be used for the annotation and study of multimodal data. In this article, I choose to focus on ELAN[1] (Sloetjes and Wittenburg 2008), a professional,

[1] ELAN can be downloaded for free at https://tla.mpi.nl/tools/tla-tools/elan/. The current version available is version 6.1.

open-source tool for the creation of complex annotations on video resources developed at the Max Planck Institute for Psycholinguistics in Nijmegen (Netherlands), which is well-suited to the annotation and study of gesture. A detailed description of the software is available at the ELAN website,[2] which I reproduce below:

> With ELAN a user can add an unlimited number of annotations to audio and/or video streams. An annotation can be a sentence, word or gloss, a comment, translation or a description of any feature observed in the media. Annotations can be created on multiple layers, called 'tiers'. Tiers can be hierarchically interconnected. An annotation can either be time-aligned to the media or it can refer to other existing annotations. The textual content of annotations is always in Unicode and the transcription is stored in an XML format.

ELAN provides different views on the annotations, and each view is connected and synchronised to the media play head. Up to four video files can be associated with an annotation document. Each video can be integrated in the main document window or displayed in its own resizable window. ELAN delegates media playback to an existing media framework, like *Windows Media Player*, *QuickTime* or *Java Media Framework* (JMF). As a result, a wide variety of audio and video formats is supported, and high-performance media playback can be achieved. ELAN is written in the *Java* programming language and the sources are available under a GPL 3 license. It runs on Windows, Mac OS X and Linux. ELAN's main other features are:

- Navigate through the media with different step sizes.
- Easy navigation through existing annotations.
- Waveform visualisation of .wav files.
- Support for template documents.
- Input methods for a variety of script systems.
- Multi-tier regular expression search, within a single document or in a selection of annotation documents.
- Support for user definable Controlled Vocabularies.
- Import and export of Shoebox/Toolbox, CHAT, Transcriber (import only), *Praat* and .csv/tab-delimited text files.
- Export to interlinear text, html, smil and subtitles text.
- Printing of the annotations.
- Multiple undo/redo.

[2] https://tla.mpi.nl/tools/tla-tools/elan/elan-description/

ELAN is an especially convenient annotation tool because it is open source, compatible with other video and audio transcription and annotation software like CLAN[3] (MacWhinney 2000) or *Praat*[4] (Boersma and Weeninck 2017), and because the annotated data can easily be exported for further analysis or statistical calculus in .csv or tab-delimited format.

2.3. What counts as "the data"

One epistemological issue raised by the study of video-recordings of political speeches is the status of the recording with respect to what counts as the data. If the data is primarily considered to be the speech as it was delivered in its original setting, in the co-presence of its addressees, then the videorecording can be seen as a mere tool for accessing the data itself. This view is, for instance, usually adopted in the field of Conversation Analysis to analyse talk-in-interaction (Mondada 2009). And yet, with the advent of media culture (television, the Internet), some recordings are arguably also designed first and foremost as mediated broadcasts. In this sense, the videorecording itself can also be regarded as primarily constitutive of the data. In a diachronic corpus like the DCPS, this second view will probably be increasingly relevant with time and with the development of mass media communication. Both views can, of course, hold true at once, with the speech being designed both for an in-person and online audience, both as an interaction and as a mediated broadcast, and, indeed, the speech as a successful interaction can contribute to its success as a broadcast.

3. EXPLOITABILITY OF THE VIDEO-RECORDING FOR MULTIMODAL ANALYSIS

For speeches given from 1950 onwards that will be included in the DCPS, video-recordings of the original delivery are retrievable online —often for free on video-sharing platforms like *YouTube*. Original videos can usually be downloaded with free online tools that convert *YouTube* videos into .mp4 video clips that can be stored on a computer or hard drive.[5] To my knowledge, there are no ethics or copyright issues

[3] http://dali.talkbank.org/clan/
[4] http://www.fon.hum.uva.nl/praat/
[5] For instance https://youtube-mp4.download/fr/free-converter, or the *Firefox Extension Downloadhelper*, https://addons.mozilla.org/fr/firefox/addon/video-downloadhelper/

related to collecting and analysing data for research purposes that are already freely available on online platforms.

Latest versions of ELAN are compatible with several video formats. Yet .mp4 is a sound choice, as a common widespread format, it can be played easily on either a Mac or a PC and, as a compressed format, it will allow for slightly lighter video files with enough quality of recording. A minimum quality of recording will be needed for hand gestures, posture changes, head movements and facial expressions to be visible to the human eye. Compressed formats can, nevertheless, impact the quality of the recordings and, in turn, hinder visual analysis. If multiple recordings of the same speech are available online, priority can be given to the less compressed format, if possible, as well as to the characteristics that will be discussed in what follows (cf. Sections 3.1–3.3.)

3.1. Camera framing

Political speeches are often filmed with a fairly close framing that leaves most of the hand gestures out (medium close-up framing; cf. Figure 1). If a recording with a larger framing is available (medium shot, from the waist up; cf. Figure 2), it can be preferred so as to capture as many gestures as possible. And yet, if only close-up framings are available, they are of course relevant for multimodal analysis as well. Indeed, political orators are often used to medium close-up framings and adapt by relying mostly on gestures that are visible on camera, such as facial expressions, head and shoulder movements, and hand gestures with a small amplitude or realised in the visible part of the gesture space.

Figure 1: Medium close-up framing

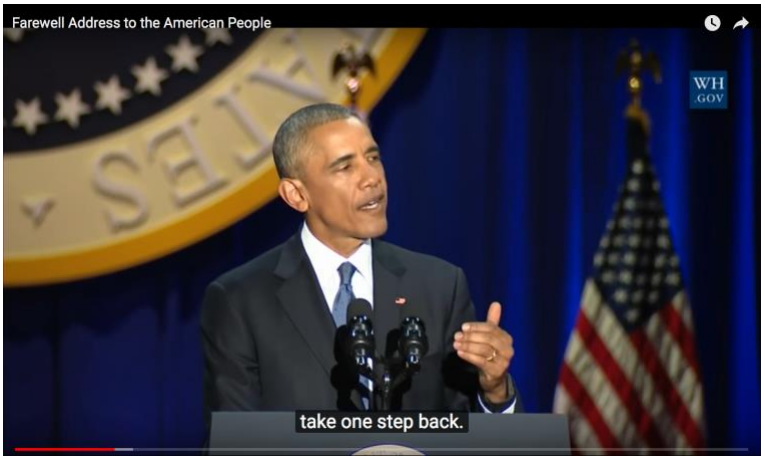

Figure 2: Medium shot framing

3.2. Camera angle

A recording that faces the speaker should be preferred to one with a camera positioned sideways. And yet, if the speaker is positioned behind a high reading desk, gestures might be partly hidden behind it. In that case, if a recording with a sideways camera angle is available, it can be preferred so as to capture more of the speaker's hand gestures (cf. Figure 3).

Figure 3: Sideways angle with high desk

3.3. Continuity of filming

Discontinuous camera shooting is frequent. Editing choices can include large camera framings of the audience or zoom-ins on certain audience members. Such interventions can result in gestures of the speaker being realised out of frame. If multiple recordings of the same speech are available, the most continuous recording should be preferred. If the most continuous (or the only available) recording potentially includes out-of-frame

gestures (e.g. framings of the audience while the speech is heard as a voice-over), the recording remains exploitable for multimodal analysis. In such (frequent) cases, the work will bear on whatever occurrences are observable.

4. TRANSCRIPTION, SEGMENTATION AND ALIGNMENT

To navigate multimodal data, the transcription of each speech must be segmented into units, and these units need to be temporally aligned with the stream of the corresponding video. When transcripts are available online, for instance on institutional websites, they are usually fairly accurate. It can happen that some aspects typical of spoken delivery have been smoothed out in written transcripts. If so, the transcript should be proofread and corrected so as to reintroduce the marks of orality (e.g. hesitations, filled pauses, discourse markers, repetitions, repairs) that are actually produced by the speaker. This would be useful so as to carry diachronic studies of the emergence of the conversational framing in political discourse (Cienki and Giansante 2014). Various units can be envisaged for the segmentation of speech. I choose to focus on two of them: the interpausal unit (automatic segmentation) and the intonation unit (manual segmentation).

4.1. The interpausal unit (IPU)

IPUs are defined, in the framework of Conversation Analysis, as blocks of speech separated by silent pauses of 0.2 seconds (Koiso *et al.* 1998). The length of the pause can vary according to the language used and to the speech situation. Since monological speech during public political address is fundamentally different from spontaneous conversation (Rossette-Crake 2019, *inter alia*), the length of separation pauses will probably need to be readjusted for some speakers in the DCPS. The automatic alignment and segmentation into IPUs of the transcript can be realised automatically, for instance, with the software *SPeech Phonetization Alignment and Syllabification* (SPPAS; Bigi 2012; Bigi and Hirst 2012).[6] SPPAS can also produce automatic annotations of word, syllable and phoneme segmentations from a recorded speech sound and its corresponding transcription. The resulting alignments are a set of *TextGrid* files, the native file format of the *Praat* software, which will need to be

[6] http://www.sppas.org/

corrected manually. They can then be converted into annotation tiers in the ELAN software, where additional gesture annotations will be made.

4.2. The intonation unit (IU)

If there is opportunity (e.g. time, funding) for finer corpus annotation work, the DCPS will be manually segmented into IUs (Chafe 1994), also known as 'intermediate phrases' (Pierrehumbert and Hirschberg 1990), 'intonation phrases' (Wells 2006), or 'tone-units' (Lelandais and Ferré 2019), which can be considered as the spoken equivalent of the clause, as theorised in the well-established tradition of the British school of intonation analysis (Halliday 1967; Wells 2006, *inter alia*). An intonation phrase is organised around at least one nucleus, and characterised by a dynamic pitch contour, and is considered to constitute an information unit (Chafe 1994).

Chafe (1994: 58) defines six characteristics of prototypical IUs, namely pitch, duration, intensity, pausing, voice quality and speaker turn. Pitch (that is, fundamental frequency) usually includes a resetting of the pitch baseline (as in a 'step- up' or 'step-down' in the pitch level) and a recognisable final pitch contour (e.g., falling or rising). Duration usually includes increased tempo at the beginning (as in a shortening of syllables and/or words), and then a gradual slowing down toward the end (as in a lengthening of syllables and/or words). Intensity usually includes one or more syllables and/or words spoken more loudly. Pausing is often preceded or followed by pausing (but may also contain pauses within its boundaries). Voice quality sometimes begins or ends with a creaky voice or whispering. Finally, speaker turn may sometimes be associated with a change of speaker.

Accordingly, the prototypical intonation phrase can be defined as follows:

> (…) it is a spate of talk delivered as one recognisable overall pitch movement. In a standard textbook scenario this pitch movement would contain a pitch accent near the beginning, and another, typically more prominent pitch accent on the final stressed syllable; it would start with a comparatively high pitch onset, which would be followed by gradual declination in overall pitch register and loudness; the last syllable would be lengthened; and the whole phrase would be followed by a brief pause (Szczepek Reed 2011: 351).

Since the annotation of intonation units needs to be done manually, notably based on acoustic analyses of pitch contours in *Praat*, inter-coder reliability will need to be

established through a statistical test, such as Cohen's Kappa (McHugh 2012), so as to ensure the robustness of the annotation. As with any form of manual annotation, differences between annotators may occur: less experienced coders can, for instance, miss some boundaries. Ideally, a section of the data will be transcribed by at least two coders: experienced and less experienced ones. That way, less experienced coders may improve their annotation skills from confronting several transcripts. I here refer the reader to Stelma and Cameron's (2007) methodological paper on building skills for Intonation Unit annotation.

5. Gesture annotations

5.1. Proposed guidelines for the formal and functional annotation of gesture

Although encouraging progress is being made in the automatisation of gesture annotation, notably thanks to motion capture technologies, most gesture annotation still has to be done manually. Gesture annotation is a time-consuming process. As explained in Section 3, a medium close-up framing is often favoured when politicians are filmed during monological public address. Therefore, the most visible gestures, which are usually also the ones that are the most mobilised by these coached public speakers, are facial expressions, head and shoulder movements, as well as hand gestures with a small amplitude or realised in the visible part of the gesture space.

Several very thorough reliable annotation systems have been developed for the annotation of gestures, such as the *Facial Action Coding System* (FACS; Ekman and Rosenberg 1997), for the annotation of facial expressions, and the *Linguistic Annotation System for Gestures* (LASG; Bressem *et al*. 2013) for the annotation of hand gestures. The main downside with these systems is precisely the direct consequence of their robustness and quality: they demand a lot of annotation time because they are extremely detailed, and since they require a significant degree of expertise and practice, they are not very adapted to beginners or non-specialists. Indeed, as Waller and Pasqualini (2013: 920) explain, the mastery of FACS is quite demanding:

> To use the system, researchers must learn to identify these base units of facial movements using a detailed manual (Ekman et al. 2002) and take a final test for certification. Although "in-house" inter-coder reliability may be desirable for specific studies in addition to

certification, it is not recommended as a substitute. Training takes an estimated 100 hours but can take more or less time depending on the context.

Likewise, LASG, which proposes the intonation unit as the unit of analysis for speech (cf. Section 4.2), constitutes a comprehensive yet quite complex analysis of the linguistic co-expressiveness of speech and gesture (McNeill 1992), as shown in Table 1.

Level of annotation		Name of Tier	obligatory/ optional	controlled vocabulary
Annotation of gestures	determining units	Gesture Unit	obligatory	
		Gesture Phases		
	annotation of form	Hand Shape	obligatory	x
		Orientation		
		Position		
		Movement Type		
		Movement Direction		
		Movement Quality		
	motivation of form	Mode of representation (MoR)	obligatory	
		Action		
		Motor pattern		
		Image schema		
Annotation of speech	annotation of speech (turn)	Speech Turn	obligatory	
		Speech Turn-translation		
		Speech Turn-Gesture Phases		
		Speech Turn-Gesture Phases translation		
	annotation of speech (intonation unit)	Intonation Unit		
		Intonation Unit-translation		
		Intonation Unit-Gesture Phases		
		Intonation Unit-Gesture Phases translation		
Annotation of gestures in relation to speech	prosody	Final pitch movement	obligatory	x
		Accent (primary, secondary)	optional	
	Syntax	Word Class	optional	x
		Syntactic Function		
		Integration		
	Semantics	Temporal Relation	obligatory	x
		Semantic Relation	optional	
		Semantic Function		
	Pragmatics	Turn	obligatory	x
		Speech Act	optional	
		Pragmatic Function		
		Dynamic Pattern		

Table 1: Overview of the levels of annotation in the LASG system (from Bressem *et al.*, 2013: 1101)

If preparatory gesture annotations can be made on the DCPS in ELAN, in the perspective of multimodal analysis, they do not need to be as fine-grained as these systems require. Depending on the annotation resources available, I suggest a series of more coarse-grained preparatory annotations that are accessible to less expert coders. These coding recommendations are rooted in a form-based approach (Boutet 2008; Müller *et al*. 2014), and partly inspired from the MUMIN annotation scheme (Allwood *et al.* 2005) and from the LASG system. Indeed, as Bressem *et al.* (2013: 1104) note:

> The Linguistic Annotation System for Gestures approaches the description of gestures' forms by applying the four parameters "hand shape", "orientation", "movement" and "position in gesture space", developed for the description of signs (Battison 1974; Stokoe 1960) to gestures. Taking the four form parameters as the basis of a gestural form description aims at systematically addressing the form aspects of a gestural Gestalt. In doing so, it allows for a fine-grained description of gestures and for a detection of gestural patterns and structures (e.g. […] Kendon 2004; Müller 2004).

If all the preparatory annotations suggested below cannot be made due to time or material constraints, we propose an order of priority to realise them. First, annotators will make formal annotations indicating what articulator is used (e.g. head, hand, eyebrows) and what configuration of the articulator (gesture form) is used, (e.g. for the head: head nod, head shake, head tilt). Second, functional annotations will be made, indicating the co-verbal function of hand gestures (McNeill 1992; Kendon 2004) and head movements (McClave 2000; Kendon 2002).

An annotation line or tier will be created in ELAN for each articulator used by the speakers. The annotation of a gesture form will include all gestures phases (cf. Kendon 2004): extension (i.e. departure from the resting or 'home' position (Sacks and Schegloff 2002), the stroke, the hold if applicable, and the retraction phase until the return to the home position, or the shift to another gesture form. For instance, in the case of a palm-up open-hand (PUOH) gesture, the annotation will start from the extension of the hand from the resting position to the gesture stroke itself during which the palm is turned upward (stroke) and possibly held in that position (gesture hold), until the hand reaches back the home position after total retraction. Gesture phases do not need to be annotated right away: their detailed annotation can be done in a dependent tier later on by a gesture researcher, if the multimodal focus of the study requires it. Each annotation can be labelled with a description of the gesture form (configuration of the body

articulator involved), according to the annotation tables proposed below. Table 2 proposes a formal annotation of movements of the shoulders (which are often connected with shrugging (see Kendon 2004; Streeck 2008; Debras 2017), and of the face, inspired from the MUMIN annotation guide (Allwood *et al*. 2005). In turn, Tables 3 and 4 propose labels for the formal and functional annotation of head movements and hand gestures respectively. Tables 2 and 3 both include references about the chosen labels in the rightmost column, should the reader wish to seek more information about them.

Annotation tier	Annotation labels
Shoulders	Both shoulders lifted
	Right shoulder lifted
	Left shoulder lifted
	Other
Mouth/ Lips	Smile
	Laughter
	Corners up
	Corners down
	Protruded
	Retracted
	Other
Gaze	Up
	Down
	Sideways
	Other
Eyes	Exaggerated Opening
	Closing-both
	Closing-one
	Closing-repeated
	Other
Eyebrows	Frowning
	Raising
	Other

Table 2: Labels for the annotation of non-manual gestures (with adaptations from Allwood *et al*. 2005)

Form (after Allwood *et al*. 2005)	Function	Corresponding reference for further information on the chosen labels
Single Nod (Down)	Assessment	Goodwin and Goodwin (1992)
Repeated Nods (Down)	Agreement	
Single Jerk (Backwards Up)	Inclusivity	
Repeated Jerks (Backwards Up)	Intensification	
Single Slow Backwards Up	Uncertainty	
Move Forward	Direct quotes	
Move Backward	Expression of mental images of characters	McClave (2000)
Single Tilt (Sideways)	Deixis and referential us of space	
Repeated Tilts (Sideways)	Lists or alternatives	
Side-turn	Lexical repair	
Shake (repeated)	Backchannelling request	
Waggle	Negation	Kendon (2002
Other	Other	

Table 3: Labels for the formal and functional annotation of head movements

	Annotation tier	Annotation labels	Corresponding reference for further information on the chosen labels
Main annotation tier for hand gestures	Hand	Both hands Left hand Right hand	
Formal and functional annotation tiers that are hierarchically dependent on the main annotation tier	Handshape (See illustrations below)	Index pointing Precision grip Vertical Palm PUOH (Palm-Up Open Hand) Thumb pointing Other	Kendon (2004) Müller (2004)
	Orientation	Vertical palm facing outwards Vertical palm facing sideways Horizontal palm down Horizontal palm sideways Horizontal palm up Oblique Other	
	Trajectory	Up Down Sideways Complex Other	Allwood *et al.* (2005)
	Localisation in the gesture space	Centre Periphery	
	Semantic relation with speech	Redundant Complementary/Supplementary Contrary Replacing	Bressem *et al.* (2013: 1111)
	Mode of representation (for iconic gestures only)	Drawing Molding Representing Acting	Müller (2014)
	Main function with respect to speech	Iconic Deictic Metaphoric Beat Emblem Interactive Recurrent gesture	McNeill (1992) Bavelas *et al.* (1992) Ladewig (2014)

Table 4: Labels for the formal and functional annotation of hand gestures

5.2. Some recurrent gesture forms in political oratory

Figures 4 to 7 present illustrations of handshapes that are recurrently used by political orators. A remarkable handshape that is used only in political oratory, by orators like Bill Clinton, Barack Obama or Justin Trudeau, is thumb pointing, also known as the 'Clinton thumb' (Mankiewicz 2006), although John F. Kennedy was first observed using it, presented in Figure 4. When they want to point to an abstract idea (abstract deixis, as per McNeill *et al.* 1993) and/or stress a point with a prosodic beat gesture (McNeill 1992), certain politicians are coached to point outward with the thumb rather

than with the index, so as to avoid the more aggressive connotations attached to finger pointing. Figure 5 shows an example of another handshape, index pointing, used with two distinct orientations.

Figure 4: Illustrations of thumb pointing

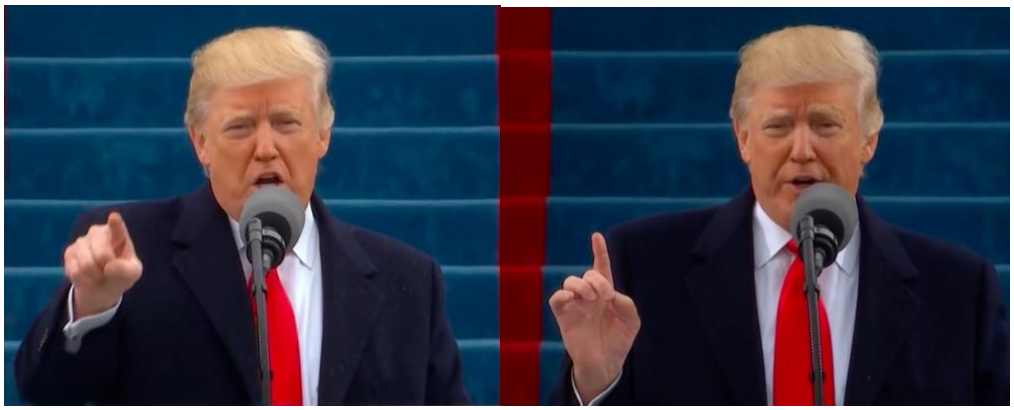

Figure 5: Index pointing with different orientations: horizontal palm down or upward

Figure 6 illustrates the ‘precision grip’ (Kendon 2004), a recurrent gesture whose core meaning is the expression of a precise, specific idea. Figure 7 is a typical example of a metaphoric gesture: the speaker is talking about an abstract referent, the topic of ‘values’, while representing this referent as a concrete object that can be manipulated — see also Streeck (2008) on speech-handling gestures.

Figure 6: Precision grip

Figure 7: Metaphoric gesture

5.3. Quantity of gesture

The multimodal component of the DCPS will provide at least several hundred kinesic forms. This is largely sufficient to conduct systematic linguistic analyses of gestures, notably of recurrent gestures and/or multimodal constructions (Steen and Turner 2013), notably with multivariate exploratory statistics tools in *R* (Desagulier 2017).

6. Conclusion

The DCPS is a collection of 1,500 full-length political speeches in English. This article aimed to present the practical and methodological issues at stake in preparing the video-recorded component of the DCPS for linguistic multimodal analysis, including the analysis of gesture, and to propose relevant recommendations and guidelines. It focused on four main issues: 1) the exploitability of recordings, 2) the segmentation and alignment of transcriptions, 3) the annotation of gesture forms and functions in ELAN and 4) the quantity of available gesture data. My main recommendations have the following:

1) Favour recordings with a medium shot framing so as to increase the visibility of the hand gestures used alongside facial displays and head movements.
2) Transcribe and segment the corpora in intonation units and rely on inter-coder agreement when segmenting the data.
3) Annotate gesture forms and functions based on the guidelines available in Tables 2, 3 and 4, which although already quite detailed are still coarser grained than existing annotation systems like FACS or LASG, and more accessible to less expert coders.

References

Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Constanza Navarretta and Patrizia Paggio. 2005. The MUMIN annotation scheme for feedback, turn management and sequencing. *Proceedings of the 2nd Nordic Conference on Multimodal Communication*. Gothenburg, Sweden. http://www.sskkii.gu.se/jems/publications/bfiles/B80-3.pdf. (28 June, 2021.)

Battison, Robin 1974. Phonological deletion in American sign language. *Sign Language Studies* 5: 1–19.

Bavelas, Janet Bavelas, Nicole Chovil, Douglas A. Lawrie and Allan Wade. 1992. Interactive gestures. *Discourse Processes* 15/4: 469–489.

Bigi, Brigitte. 2012. SPPAS: A tool for the phonetic segmentation of speech. 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, 1748–1755.

Bigi, Brigitte and Daniel Hirst. 2012. *SPeech Phonetization Alignement and Syllabification* (SPPAS): A tool for the automatic analysis of speech prosody. In Qiuwu Ma, Hongwei Ding and Daniel Hirst eds. *Proceedings of 6th Speech Prosody International Conference*. Shanghai: Tongji University Press, 19–22.

Boersma, Paul and David Weenink. 2017. *Praat: Doing Phonetics by Computer*. Computer Program. Version 6.0.28.

Boutet, Dominique. 2008. Une morphologie de la gestualité: Structuration articulaire. *Cahiers de Linguistique Analogique* 5: 81–115.

Bressem, Jana, Silva H. Ladewig and Cornelia Müller. 2013. Linguistic Annotation System for Gestures (LASG). In Cornelia Müller *et al*. eds., 1098–1125.

Chafe, Wallace. 1994. *Discourse, Consciousness and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. Chicago: University of Chicago Press.

Cienki, Alan. 2009. Spoken language framing in political discourse. Presentation at the European Consortium for Political Research (ECPR). *Workshop Studying the Political through Frame Analysis*, 14–19 April 2009. Lisbon: Portugal.

Cienki, Alan. 2017. Language as a prototype category. In Alan Cienki ed. *Ten Lectures on Spoken Language and Gesture from the Perspective of Cognitive Linguistics: Issues of Dynamicity and Multimodality*. Leiden: Brill, 163–182.

Cienki, Alan and Gianluca Giansante. 2014. Conversational framing in televised political discourse: A comparison from the 2008 elections in the United States and Italy. *Journal of Language and Politics* 13/2: 255–288.

Debras, Camille. 2017. The shrug: Forms and functions of a compound enactment. *Gesture* 16/1: 1–34.

Debras, Camille and Émilie L'Hôte. 2015. Framing, metaphor and dialogue – A multimodal approach to party conference speeches. *Metaphor and the Social World* 5/2: 177–204.

Desagulier, Guillaume. 2017. *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics*. New York: Springer.

Ekman, Paul and Erika Rosenberg eds. 1997. *What the Face Reveals. Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford: Oxford University Press.

Ekman, Paul, Wallace V. Friesen and Joseph C. Hager. 2002. *Facial Action Coding System*. Salt Lake City: Research Nexus.

Ferré, Gaëlle. 2011. Functions of three open-palm hand gestures. *Multimodal Communication* 1/1: 5–20.

Ferré, Gaëlle. 2019. *Analyse de Discours Multimodale. Gestualité et Prosodie en Discours*. Grenoble: Éditions de l'Université Grenoble Alpes.

Goodwin, Charles and Marjorie Harness Goodwin. 1992. Assessments and the construction of context. In Alessandro Duranti and Charles Goodwin eds. *Rethinking Context*. Cambridge: Cambridge University Press, 147–190.

Heritage, John and David Greatbatch. 1986. Generating applause: A study of rhetoric and response at party political conferences. *American Journal of Sociology* 92/1: 110–157.

Halliday, Michael Alexander Kirkwood. 1967. *Intonation and Grammar in British English*. The Hague: Mouton de Gruyter.

Kendon, Adam. 1986. Some reasons for studying gesture. *Semiotica* 62/1–2: 3–28.

Kendon, Adam. 2000. Language and gesture: Unity or duality? In David McNeill ed., *Language and Gesture*. Cambridge: Cambridge University Press, 47–63.

Kendon, Adam. 2002. Some uses of the head shake. *Gesture* 2/2: 147–182.

Kendon, Adam. 2004. *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press.

Koiso, Hanae, Yasuo Horiuchi, Akira Ichikawa and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech* 41/3–4: 295–321.

Ladewig, Silva H. 2014. Recurrent gestures. In Cornelia Müller *et al*. eds., 1558–1574.

Lelandais, Manon and Gaëlle Ferré. 2019. The verbal, vocal and gestural expression of (in)dependency in two types of subordinate constructions. *Journal of Corpora and Discourse Studies* 2: 117–143.

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk.* Mahwah: Lawrence Erlbaum Associates.

Mankiewicz, Josh. 2006. *For Politicians, the Gesture's the Thing*. https://www.nbcnews.com/id/wbna15609023. (28 June, 2021.)

McClave, Evelyn. 2000. Linguistic functions of head movements in the context of speech. *Journal of Pragmatics* 32/7: 855–878.

McHugh, Mary L. 2012. Interrater reliability: The kappa statistic. *Biochemia Medica* 22/3: 276–282.

McNeill, David. 1992. *Hand and Mind. What Gestures Reveal about Thought*. Chicago: University of Chicago Press.

McNeill, David, Justine Cassell and Elena Levy. 1993. Abstract deixis. *Semiotica* 1/2: 5–19.

Mondada, Lorenza. 2009. Video recording practices and the reflexive constitution of the interactional order: some systematic uses of the split-screen technique. *Human Studies* 32/1: 67–99.

Mondada, Lorenza. 2016. Multimodal resources and the organization of social interaction. In Andrea Rocci and Louis de Saussure eds. *Verbal Communication*. Berlin: De Gruyter, 329–350.

Müller, Cornelia. 2004. Forms and uses of the palm up open hand: A case of a gesture family? In Cornelia Müller and Roland Posner eds., *The Semantics and Pragmatics of Everyday Gestures*. Berlin: Weidler Buchverlag, 233–256.

Müller, Cornelia. 2014. Gestural modes of representation as techniques of depiction. In Cornelia Müller *et al*. eds., 1687–1701.

Müller, Cornelia, Jana Bressem and Silva H. Ladewig. 2014. Towards a grammar of gestures: A form-based view. In Cornelia Müller *et al*. eds., 707–732.

Müller, Cornelia, Silva H. Ladewig and Jana Bressem. 2013. Gesture and speech from a linguistic perspective: A new field and its history. In Cornelia Müller *et al.* eds., 55–81.

Müller, Cornelia, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Sedinha Teßendorf eds. 2013. *Body – Language – Communication Vol. 1*. Berlin: Walter de Gruyter.

Müller, Cornelia, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Sedinha Teßendorf eds. 2014. *Body – Language – Communication Vol. 2*. Berlin: Walter de Gruyter.

Norris, Sigrid. 2004. *Analyzing Multimodal Interaction: A Methodological Framework*. New York: Routledge.

Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In Philip R. Cohen, Jerry Morgan and Martha E. Pollack eds. *Intentions in Communication*. Cambridge: MIT Press, 271–311.

Rossette-Crake, Fiona. 2019. *Public Speaking and the New Oratory: A Guide for Non-native Speakers*. London: Palgrave Macmillan.

Sacks, Harvey and Emanuel A. Schegloff. 2002. Home position. *Gesture* 2/2: 133–146.

Schegloff, Emanuel A. 1984. On some gestures' relation to talk. In J. Maxwell Atkinson and John Heritage eds. *Structures of Social Action*. Cambridge: Cambridge University Press, 266–296.
Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, United Kingdom. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf. (28 June, 2021.)
Sloetjes, Hans and Peter Wittenburg. 2008. Annotation by category – ELAN and ISO DCR. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk and Daniel Tapias eds. *Proceedings of the 6th International Conference on Language Resources and Evaluation*, LREC 2008. Marrakech: European Language Resources Association, 816–820.
Stam, Gale and Mika Ishino eds. 2011. *Integrating Gestures: The Interdisciplinary Nature of Gesture*. Amsterdam: John Benjamins.
Steen, Francis and Mark Turner. 2013. Multimodal Construction Grammar. In Mike Borkent, Barbara Dancygier and Jennifer Hinnell eds. *Language and the Creative Mind*. Chicago: University of Chicago Press, 255–274.
Stelma, Juurd H. and Lynne J. Cameron. 2007. Intonation units in spoken interaction: Developing transcription skills. *Text and Talk* 27/3: 361–393.
Stokoe, William. 1960. *Sign Language Structure*. Buffalo: Buffalo University Press.
Streeck, Jürgen. 2008. Gesture in political communication: A case study of the democratic presidential candidates during the 2004 primary campaign. *Research on Language and Social Interaction* 41/2: 154–186.
Szczepek Reed, Beatrice. 2011. Units of interaction: 'Intonation phrases' or 'turn constructional phrases'? In Hi-Yon Yoo and Élisabeth Delais-Roussarie eds. *Proceedings of the International Conference on Prosody-Discourse Interface*, 351–363.
Waller, Bridget and Marcia Pasqualini. 2013. Analysing facial expression using Facial Action Coding Systems (FACS). In Cornelia Müller *et al*. eds., 917–931.
Wehling, Elisabeth. 2009. Argument is gesture war. Function, form and prosody of discourse structuring gestures in political argument, *Proceedings of the 35th Annual Meeting of the Berkeley Linguistics Society*: 54–65.
Wells, John C. 2006. *English Intonation: An Introduction*. Cambridge: Cambridge University Press.
Zima, Elisabeth and Alexander Bergs eds. 2017. Special issue: Towards a multimodal construction grammar. *Linguistic Vanguard* 3/1.

Corresponding author
Camille Debras
Centre de Recherches Anglophones
200 Avenue de la République
Université Paris Nanterre
92001 Nanterre Cedex
France
Email: cdebras@parisnanterre.fr

received: March 2020
accepted: June 2021

RiCL Research in Corpus Linguistics

Review of Fuster-Márquez, Miguel, Carmen Gregori-Signes and José Santaemilia Ruiz eds. 2020. *Multiperspectives in Analysis and Corpus Design*. Granada: Comares. ISBN: 978-84-1369-009-4

Moisés Almela-Sánchez
University of Murcia / Spain

The growth of a discipline is usually welcomed by the specialised academic community, but success is quickly followed by new challenges, and accomplishment gives way to the difficult task of defining new goals. This task is often a source of controversy, because setting new goals may involve redefining boundaries. As the research scope of the discipline is expanded, its limits with neighbouring disciplines are blurred, and old debates about the genuine aims and foundational principles of the discipline may be reignited.

The evolution of corpus linguistics provides a good illustration of this process. The debate about the nature of corpus linguistics and the different ways of approaching its definition dates back to earlier stages of the discipline (see Leech 1992), but the question took on a new dimension at the turn of the century as corpus methods came to be incorporated in studies from an ever wider diversity of theoretical backgrounds (including cognitive and structural linguistics, among others), and disciplines which had remained remote from a corpus linguistic approach to language, such as psycholinguistics, turned more and more frequently to corpus research in search of triangulated evidence. This proliferation of roles attributed to corpus evidence has not been free of controversy, as different influential voices in the field hold diverging views on whether certain ways of using corpora are more genuine than others. The debate between conflicting versions of corpus linguistics was particularly intensive —and proportionally fertile— in the first decade of the new century, and the relationship between theory and methodology was soon established as a central issue in the discussion (see, among others, Tognini-Bonelli 2001; Meyer 2002; Teubert 2005; Parodi 2008; Gries 2010).

Research in Corpus Linguistics 9/1: 152–159 (2021). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.09.01.09

It is plausible to affirm that, over the last decade, the more expansive definitions of corpus linguistics have taken the lead. The idea of a privileged bond between corpus linguistics and particular linguistic traditions or theoretical approaches has waned in recent years, and the field has accelerated the pace of its advances in a multiplicity of directions. Today, corpus linguistics is predominantly regarded as a framework of methodological resources compatible with, and valuable to, diverse paradigms of linguistic research and areas of inter-disciplinary exchange.

The edited collection under review is an eloquent testimony to this rich diversity. The selection of papers in the volume gives concise expression to the multiplicity of perspectives and approaches that have fed the growth of corpus research and stimulated its spread across disciplinary boundaries. The volume has relatively compact dimensions. It consists of nine contributions occupying a space of less than 130 pages, a size which is not larger than average among edited volumes. Remarkably, within these compact dimensions, the editors have managed to fit a collection of papers which represent diverse areas of research, both theoretical and applied, and which serve to illustrate some of the key trends observed in contemporary corpus linguistics. Thus, the volume strikes a difficult balance between comprehensiveness and focus. The collection is both succinct and informative. In a condensed manner, it conveys a sense of the polyvalent character of corpus methods, and it shows how they can be adapted to meet the needs of varying and highly specific research demands.

The volume covers topics in various areas of linguistic research (historical linguistics, sociolinguistics, pragmatics, discourse analysis, specialised languages, translation), but there is a common thread running through the diverse parts. All the contributions contained in the collection exploit the flexibility of corpus tools and show how they can be adapted to suit the particular needs of highly specific research goals. There are three main ways in which this strategy is implemented in the contributions contained in the volume. In most of them, the authors have compiled a corpus which is specifically designed for a particular research purpose or project. This is the case of the chapters authored by Arinas Pellón and Anesa (pp. 1–13), Pérez Ruiz and Ortego Antón (pp. 15–31), Verdaguer, Castaño and Laso (pp. 62–72), Serrat Roozen (pp. 73–88), Moreno-Sandoval, Gisbert and Montoro (pp. 89–102), and Vázquez García and Férnandez-Montraveta (pp. 115–127). In other studies, the authors take full advantage of the internal structure of existing corpora. The contributions by Rodríguez-Abruñeiras (pp.

33–45) and Tamaredo (pp. 47–60) are paradigmatic examples of how to exploit the potential of subcorpora divisions for conducting comparisons of multiple descriptive variables. Finally, the contribution by Romero-Barranco (pp. 103–114) represents a third way of exploiting the versatility of corpus tools, since it highlights the possibility to adapt the use of particular tools to heterogeneous types of corpora. In particular, he shows that corpus tools which were originally designed to process Present-day English can also be employed in historical linguistics, provided the appropriate techniques are applied.

As befits a volume on corpus linguistics, all the contributions devote substantial attention to the description of methodological aspects. In some chapters, this special emphasis includes a detailed account of the criteria applied in the design of a specially created corpus. In other chapters, the emphasis on methodological aspects takes a different form, with a focus on the process of corpus annotation, on the adaptation of part-of-speech tagging tools, or on the selection of subcorpora. Overall, the collection highlights the potential of the corpus linguistic methodological framework for providing tailor-made solutions to highly specific research objectives.

The volume opens with an introduction by the editors, as is customary in this type of collections, followed by the chapter "Advanced-fee scams: A corpus and genre analysis" by Ismael Arinas Pellón and Patrizia Anesa. This paper analyses the language used in scam emails. The data are extracted from the *Corpus of Advanced-Fee Scams* (CAFS), a corpus consisting of more than 500 emails. The analytical framework is multidisciplinary, as it combines insights from neo-Firthian linguistics, genre analysis, and psychology. The identification of linguistic patterns is based on the classical Sinclairian model of extended lexical analysis —expounded also by Stubbs (2002)— with its distinction of four main descriptive categories: collocation, colligation, semantic preference, and semantic prosody. The patterns detected in the corpus are then related to categories of motivational choices and persuasion strategies. One of the most interesting conclusions from the study is that scam emails can be analysed as a variant of sales promotion letters, since they contain similar rhetorical moves, offer similar types of incentives to the recipients and use similar strategies to generate credibility. As the authors point out, research of this type, which identifies patterns of language use in fraudulent emails, can contribute to the development of systems capable of detecting and neutralising these attempts. Another potential application of this type of research is to help educate and alert the public about the typical characteristics of scam emails.

The second paper is “El sabor de las manzanas: análisis contrastivo (español-inglés) de la terminología objetiva referida a la experiencia sensorial del gusto” by Leonor Pérez Ruiz and María Teresa Ortego Antón. The language patterns analysed in this study correspond to the description of gustatory perceptions. The data are obtained from two comparable corpora (in English and Spanish, respectively) consisting of fact sheets on apples gathered from websites of food companies. The results from the study highlight the richness of the terminology employed to describe gustatory sensations. The conclusions also indicate that these descriptions tend to focus on four main aspects, namely, 1) the degree of sweetness/acidity, 2) the evocation of other types of food and beverage, 3) the aroma, and 4) the touch, and that they are often accompanied by lexical intensifiers and downtoners which help to convey subtle nuances. The study points to potential applications in the marketing strategies used by food companies.

The third contribution is “Two example markers in and beyond exemplification: Dialectal, register and pragmatic considerations in the 21st century” by Paula Rodríguez-Abruñeiras. This study provides a thorough analysis of the use of two example markers (*for example* and *for instance*) in two corpora representing different geographical varieties of English: *British English 2006* (BrE06) and *American English 2006* (AmE06). The author applies a threefold typology of the uses of exemplary markers —exemplification, selection, argumentation— and analyses the distribution of these uses in different text types of the two corpora. This serves to take into account the interplay of register and dialectical variables. The analysis is further enriched with the consideration of different positions occupied by example markers (before their scope domain, after their scope domain, and in the middle of the example) and an analysis of their effects on the pragmatic functions. The results indicate that different positions tend to be associated with different pragmatic nuances, such as focus or mitigation. In sum, the study provides a valuable contribution to the analysis of discourse markers in English, since it offers a highly systematic and fine-grained description and takes various relevant aspects into account (dialect, register, position).

The study of language variation is also at the heart of the next contribution “Probabilistic grammars across registers: Pronominal subject expression in some varieties of English” by Iván Tamaredo. This paper investigates which factors, both language-internal and language-external, act as the most effective determinants of the choice between overt and omitted pronominal subjects. The data analysed are obtained from

three components (British, Indian, Singaporean) of the *International Corpus of English* (ICE), and the analytical framework combines elements of probabilistic grammar and of research into World Englishes. Following a sophisticated quantitative and qualitative analysis, the author concludes that clause position and coordination are the most important language-internal constraints on the distribution of pronoun omission across varieties, modes of production, and levels of formality, and that mode of production and level of formality are the most powerful language-external factors. This paper is remarkable for its methodological rigour and depth of analysis.

The next contribution, entitled "Semantic frames in *SciE-Lex*" (Isabel Verdaguer, Emilia Castaño and Natalia Judith Laso), presents recent advances in a specialised lexicographic resource. *SciE-Lex* is a lexical database of biomedical English developed by the *GreLic Research Group* at the University of Barcelona.[1] The empirical data for this database are obtained from the *Health Science Corpus* (HSC), compiled by the same research group. In the current stage of development of this lexicographic project, the database is being enriched with information about semantic structures above the level of the individual lexical items. This will be useful for integrating the description of words that share a semantic background. The theoretical model applied is informed by the Fillmorean notion of 'semantic frame'. This paper is thus a good example of how a corpus linguistic methodology can be combined with a theoretical framework informed by cognitive linguistics. The proposal is illustrated with the analysis of two verbs, *to block* and *to inhibit*, which in the *Health Science Corpus* are used to evoke the frame 'Hindering'. The results of the analysis highlight the specific properties of this frame in biomedical English, compared to its description for general English in *FrameNet*. As the authors explain, the results obtained from this type of research can be used to assist dictionary users in their scientific writing.

The title of the sixth contribution in the volume is "Accesibilidad, traducción audiovisual y normas en la subtitulación online: EMPAC (*EuroparlTV Multimedia Parallel Corpus*)" by Iris Serrat Roozen. The goal of this paper is to find out whether the subtitling of the online television channel EuroparlTV conforms to the norms of audiovisual translation commonly accepted in more traditional media (TV, DVD, cinema, etc.). The corpus compiled for this purpose is the *EuroparlTV Multimedia Parallel Corpus* (EMPAC), consisting of audiovisual documents hosted in the aforementioned

[1] http://www.ub.edu/grelic/eng/?page_id=13

television channel. In particular, the study focuses on the analysis of four features related to reading speed —characters per second, characters per line, pauses between subtitles, and segmentation— and it sets out to determine whether they comply with standard recommendations. The conclusion is that, in general, they do not follow such norms, although the extent to which they deviate from them shows variations depending on the year and on the particular feature under scrutiny. The author discusses implications for the accessibility of online content.

The compilation of a specialised financial corpus is the focus of the next contribution: "*FinT-esp*: A corpus of financial reports in Spanish" by Antonio Moreno-Sandoval, Ana Gisbert and Helena Montoro. The paper provides a detailed description of the steps taken in the process of creating a corpus of Spanish financial narratives. The corpus (*FinT-esp*) consists of annual reports and financial statements published on corporate websites of companies listed in the *Madrid Stock Exchange* for the 2014–2017 period. Additionally, the authors explain the reasons for creating a more specific corpus consisting of letters to shareholders, which constitute a particularly relevant section in annual reports. A further distinction is made between two subcorpora consisting of letters to shareholders written by Presidents and by CEOs, respectively (these are expected to articulate different types of narrative). The paper offers a meticulous justification for the decisions made in the design of the corpus, and it illustrates how this resource can facilitate the application of corpus linguistic and computational techniques to analyse financial texts in Spanish.

The contribution by Jesús Romero-Barranco addresses a problem which specifically affects the creation and analysis of historical corpora. The title of this chapter is "Spelling normalisation and POS-tagging of historical corpora: The case of GUL, *MS Hunter 135* (ff. 34r-121v)." The paper highlights the benefits that the normalisation of spelling can offer for POS-tagging. This is illustrated with the processing of a specific manuscript: *MS Hunter 135*, a medical volume written in the first half of the sixteenth century. The tool for normalising spelling which is applied in this study is VARD, developed at the University of Lancaster, and the POS-tagging system is CLAWS. The results indicate that the accuracy of this POS-tagger for specific parts of the *MS Hunter 135* text can be increased by approximately 15 per cent if spelling is normalised. Based on these results, the author argues that tools which were originally designed to process

Present-day English can be adapted to historical corpora if they are complemented by appropriate strategies.

The collection ends with the chapter “Annotating factuality in the TAGFACT corpus” by Glòria Vázquez García and Ana Fernández-Montraveta. This contribution provides a detailed account of the annotation scheme devised in the TAGFACT project. The aim of this project is to create an automatic tool for the annotation of factuality, i.e. the degree of certainty with which situations are presented in texts. In principle, the tool has been created for the annotation of a Spanish corpus, but the authors argue that it can also be applied to other languages. The paper explains the criteria used for selecting the predicates to be annotated and the type of linguistic clues employed to establish the factual status. Another important aspect which receives special attention from the authors is the classification of situations into dynamic and non-dynamic ones. The authors underline the innovative character of their contribution by remarking that there is no other resource with equivalent characteristics for Spanish.

Through this diversity of topics, lines of research and applications, the selection of papers covered in the volume will give the reader an accurate portrayal of one of the key aspects that is marking the evolution of contemporary corpus linguistics, namely its tendency to cross the traditional boundaries of the discipline and to be diversified with the incorporation of a broad range of linguistic paradigms and inter-disciplinary exchanges. This does not mean that the idea of corpus linguistics as a theoretically specific and relatively homogeneous field, defined by a close connection with a particular linguistic tradition, has been completely abandoned. In fact, a substantial amount of the corpus linguistic literature produced today has a clear neo-Firthian background. However, the broader approach to the concept of corpus linguistics has been gaining ground in recent years. The number of scholars undertaking corpus research from diverse perspectives has been increasing in the last decade, and this has contributed to highlighting the potential of corpora as a pool of methodological resources compatible with multiple theories and paradigms. The volume reviewed here is a reflection of this trend and, therefore, it will be useful for readers who want to keep up-to-date with developments in the field.

REFERENCES

Gries, Stefan Th. 2010. Corpus linguistics and theoretical linguistics. A love–hate relationship? Not necessarily... *International Journal of Corpus Linguistics* 15/3: 327–343.

Leech, Geoffrey. 1992. Corpora and theories of linguistic performance. In Jan Svartvik ed. *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter, 105–122.

Meyer, Charles F. 2002. *English Corpus Linguistics*: *An Introduction*. Cambridge: Cambridge University Press.

Parodi, Giovanni. 2008. Lingüística de corpus: una introducción al ámbito. *Revista de Lingüística Teórica y Aplicada* 46/1: 93–119.

Stubbs, Michael. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Teubert, Wolfgang. 2005. My version of corpus linguistics. *International Journal of Corpus Linguistics* 10/1: 1–13.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Reviewed by

Moisés Almela-Sánchez
Departamento de Filología Inglesa
Facultad de Letras
C/ Santo Cristo s/n
30011 Murcia
e-mail: moisesal@um.es