palgrave pivot

Building a National Corpus

A Welsh Language Case Study

Dawn Knight · Steve Morris · Laura Arman · Jennifer Needs · Mair Rees

palgrave macmillan

Building a National Corpus

"The Welsh National Corpus is an important state-of-the-art corpus. Written with great precision and honesty, this book brings the reader inside the anatomy of the corpus and in doing so creates both an important archive of decision-making and protocols. Cutting-edge insights are also presented through case studies on spoken, written and e-language data from the corpus. For anyone attempting to build a corpus, this book is essential reading."

—Anne O'Keeffe, *University of Limerick, Ireland*

Dawn Knight · Steve Morris · Laura Arman ·
Jennifer Needs · Mair Rees

Building a National Corpus

A Welsh Language Case Study

palgrave
macmillan

Dawn Knight
School of English
Communication & Philosophy
Cardiff University
Cardiff, UK

Laura Arman
Wales Institute of Social and
Economic Research and Data
Cardiff University
Cardiff, UK

Mair Rees
Newport, UK

Steve Morris
School of Culture and
Communication
Swansea University
Swansea, UK

Jennifer Needs
Cardiff, UK

ISBN 978-3-030-81857-9 ISBN 978-3-030-81858-6 (eBook)
https://doi.org/10.1007/978-3-030-81858-6

Cover illustration: © Harvey Loake

This Palgrave Macmillan imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

It is estimated that half of the world's population speak one of twenty-three languages (Eberhard et al., 2020). These twenty-three languages, which include, for example, English and Spanish, are generally well-resourced. They possess, among other resources, substantial and varied language corpora. These corpora are used to inform corpus and acquisition planning (concepts typically used in language planning research, e.g. Cooper, 2010), pedagogy and language technology development, enabling speakers of these languages to, for example, engage with their personal devices using their own voices, command their television sets to search for specific programmes or enable their PCs to transcribe their speech into written form. This book focuses on building national corpora mainly (but not exclusively) for those languages which are spoken by the other 50% of the world's population.

This book's prime focus is on how a circa 11-million-word corpus of contemporary Welsh was built and how this process can help inform those working with other languages to apply/adapt the methodology to their own needs. The book provides a micro-level, working model of a methodological approach to, and practical guidelines for, building a corpus informed by the work on the CorCenCC project (Corpws Cenedlaethol Cymraeg Cyfoes—The National Corpus of Contemporary Welsh). It is hoped that through sharing this approach to corpus building in Welsh and some of the unique challenges faced, others might be encouraged to attempt similar projects to develop resources in, and maintain the health and vitality of, their own language contexts and communities.

There is not one overarching template for building a corpus that will fit all languages. Similarly, the resources and approaches used to create corpora in larger and/or well-resourced languages are often not available or wholly relevant to corpus building in what might be termed under-resourced, minority, minoritised, non-dominant or lesser-used languages. Nevertheless, the path to creating CorCenCC is one that others may wish to follow or adapt to meet the challenges of embarking on similar corpus building projects for their own language communities.

References

Cooper, R. L. (2010). *Language planning and social change*. Cambridge University Press.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (2020). *Ethnologue: Languages of the World. Twenty-third edition [Online]*. Retrieved from http://www.ethnologue.com. Accessed 20 June 2021.

Acknowledgements

The research on which this book is based was funded by the UK Economic and Social Research Council (ESRC) and Arts and Humanities Research Council (AHRC) as part of the *Corpws Cenedlaethol Cymraeg Cyfoes (The National Corpus of Contemporary Welsh): A community driven approach to linguistic corpus construction* project (Grant Number ES/M011348/1). Thanks must also be given to former members and volunteers on the project's Work Package 1 (WP1) team who helped to collect and collate the data in CorCenCC, including: Laurence Anthony, Jeremy Evas, Tesni Galvin, Alys Greene, Sioned Johnson-Dowdeswell, Alex Lovell, Jonathan Morris, Mark Stonelake, Bethan Tovey-Walsh, Gareth Watkins, Lowri Williams, Lowri Williams, Katharine Young. The authors of this book would also like to acknowledge the extended CorCenCC team and its Co-Investigators (CIs), Research Assistants (RAs), consultants, stakeholders, ambassadors, as well as all data contributors and transcribers, without whom the project, and building the corpus, would not have been possible.

Contents

List of Figures

List of Tables

CHAPTER 1

Understanding the Language Context

Abstract The book aims to provide a micro-level, working model of a methodological approach to, and practical guidelines for, building a corpus informed by the work on the CorCenCC project. This first chapter provides the context to the work and lays foundations for the first step of corpus building, that is, the examination of the context of the language. It begins with an exploration of what can be considered distinct about a 'national' corpus, and then examines how context affects the nature of a national corpus, including the differences between national corpora of major languages (both well-resourced and under-resourced) and those in under-resourced and/or minoritised languages. Following these discussions, the chapter overviews some existing language corpora created in under-resourced languages and explains the status of the Welsh language within this landscape.

Keywords Under-resourced language · Minority language · Minoritised language · Non-dominant language · Lesser-used language · National corpora · Sociolinguistic audit · Language context

D. Knight et al., *Building a National Corpus*,
https://doi.org/10.1007/978-3-030-81858-6_1

Overview and Aims

Given the increasing importance, wide utility and applications of corpora and corpus-based methods, there is a growing realisation that the creation of language resources, including corpora and corpus tools, is vital in sustaining and developing what are often described as under-resourced, minority, minoritised, non-dominant or lesser-used languages (i.e. typically those languages outside of the twenty-three spoken by half the world's population referred to in the foreword). The European Language Resources Association (ELRA—www.elra.info/en), for example, established a workshop series dedicated to Less-Resourced Languages (LRL), as well as a Special Interest Group for Under-resourced Languages (SIGUL), in collaboration with the International Speech Communication Association (ISCA—www.isca-speech.org). There are also an increasing number of research projects (including the Digital Language Diversity Project—www.dldp.eu) and academic conferences being held that focus on this topic, and numerous academic articles have been published on this in a range of computational, engineering, linguistics and other journals (e.g.Besacier et al., 2014; El-Haj et al., 2015; Scannell, 2007).

This book focuses on building national corpora mainly (but not exclusively) for under-resourced languages. It discusses, specifically, the creation of a corpus in the Welsh-language context: CorCenCC (*Corpws Cenedlaethol Cymraeg Cyfoes*—the National Corpus of Contemporary Welsh). Although Welsh is the language under discussion here, the aim is that sharing the experience of building CorCenCC through this book will enable many of the processes and approaches discussed to be applied to a lesser or greater extent to other language contexts. Wales is a bilingual country where the 2011 census (Welsh Government, 2012) recorded 562,016 speakers of Welsh (19% of the population of Wales) with varying geographical densities, age profiles, acquisition trajectories and reading and writing skills (ONS, 2011). CorCenCC is the first large-scale contemporary corpus of Welsh to account for these variations, capturing language usage across communication modes (spoken, written, e-language), genres, language varieties (regional and social) and contexts.

CorCenCC was 'user-driven' (Knight et al., 2021: 44–53) in its design, engaging closely with contemporary users of Welsh in the composition and promotion of the resource. This engagement with the users of Welsh was fundamental in ensuring that the end product was relevant to the language community and accurately reflected its contemporary constituency. The corpus, which was released in November 2020 (available at: www.corcencc.org), provides a resource to enhance and facilitate corpus and acquisition planning in Wales through, for example, supporting language technology (speech recognition, predictive text, etc.) and lexicography as well as pedagogical and academic research.

A description of the main objectives and achievements of the wider CorCenCC project is available in the final project report (Knight et al., 2020). Building on this, Knight et al. 2021 provide a rationale for, and realisation of the wider CorCenCC project (across all areas of activity). They present the conceptual framework for the user-driven design that underpinned the project, offering a blueprint for other researchers embarking on projects of this nature.

Complementing Knight et al. 2021's *conceptual framework*, this book is designed to be of significant value and relevance to those specifically interested in critically engaging with corpus building methodology. The book aims to provide a more *micro-level, working model of a methodological approach to, and practical guidelines for, building a corpus* informed by the work on the CorCenCC project. It focuses on the development of:

i. detailed design frames for corpora across communicative modes (spoken, written and e-language), and
ii. the practical processes involved in the planning, collection, transcription, collation and (re)presentation of language data.

It will become clear in the book that, to build a corpus, certain key elements need to be in place. In the case of CorCenCC, the project focused on five Work Packages (WPs) and organised the work through these. The WPs were interlinked and interdependent, collectively contributing to the realisation of the project's aims. The project team consisted of experts from corpus and applied linguistics, Welsh-language linguistics, pedagogy and computer science, from institutions in Wales, the rest of the UK and beyond. WP2 focused on the creation of a Welsh-language part-of-speech (POS) tagger and tagset; WP3 on a semantic tagger and tagset; WP4 on the development of a pedagogic toolkit and

WP5 on the creation of the corpus query tools and infrastructure for housing CorCenCC.

Reflections of the work carried out on WP1 are central to this book. The main objective of WP1 was to assemble a 10-million-word Welsh-language dataset (although as shown later, a figure of over 11-million-words was achieved). The key challenges/tasks associated with this objective were to:

1. create an appropriate design frame for the corpus
2. recruit potential participants/sources of data
3. collect written, spoken and e-language texts
4. design and apply transcription and anonymisation protocols to the data.

This book unpacks each of the challenges/tasks. It is our intention, through outlining the processes followed, and the decisions made in relation to each of these challenges/tasks, to provide guidance and a resource to other corpus and applied linguists who may be interested in building their own language corpus. To this end, this book is designed to provide:

i. a working model, and
ii. an account of building a corpus dataset, from which step-by-step guidelines for creating other linguistic corpora in *any* language can be easily extrapolated.

This first chapter provides the context to the work and lays foundations for the first step of corpus building, that is, the examination of the context of the language (using Welsh as a worked example of some of the things that need to be considered). It begins with an exploration of what can be considered distinct about a 'national' corpus, and then examines how context affects the nature of a national corpus, including the differences between national corpora of major languages (both well-resourced and under-resourced) and those in under-resourced and/or minoritised languages. Following these discussions, the chapter overviews some existing language corpora created in under-resourced languages and explains the status of the Welsh language within this landscape. Chapters 2, 3 and 4 detail some key stages of corpus building, illustrated in Fig. 1.1. This includes a focus on creating corpus design frames

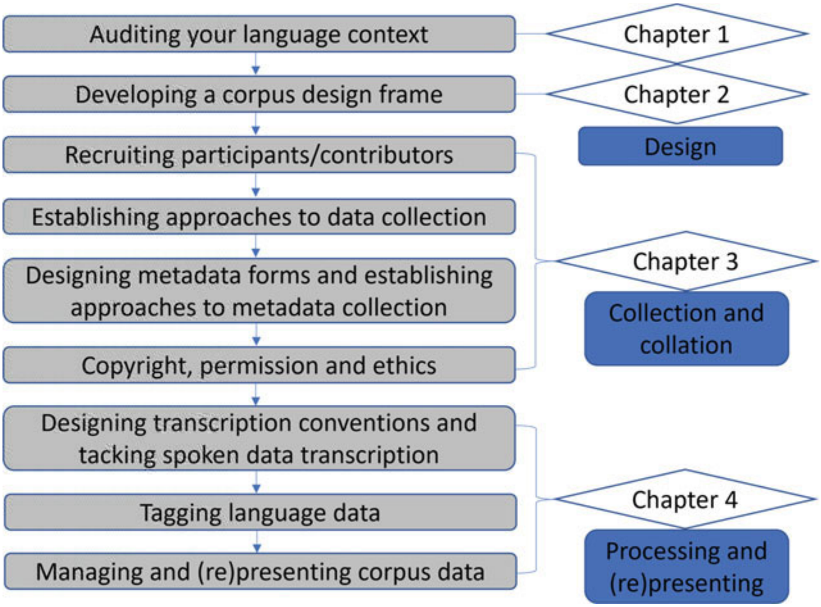
Auditing your language context
Developing a corpus design frame
Recruiting participants/contributors
Establishing approaches to data collection
Designing metadata forms and establishing approaches to metadata collection
Copyright, permission and ethics
Designing transcription conventions and tacking spoken data transcription
Tagging language data
Managing and (re)presenting corpus data
Chapter 1
Chapter 2
Design
Chapter 3
Collection and collation
Chapter 4
Processing and (re)presenting

Fig. 1.1 Key stages of corpus building explored in this book

(Chapter 2), (meta)data collection and collation (Chapter 3) and data processing and (re)presentation (Chapter 4). While these chapters relate to the experiences of building CorCenCC, they aim to inform corpus building in other language contexts (including applicability to major languages).

Chapter 5 then examines to what extent the original plans for CorCenCC were realised and offers some reflections on the challenges faced when building the corpus, with practical advice on overcoming these.

National Corpora

So, what is a corpus? Broadly defined, a corpus (plural = corpora) is an electronic database of words. Corpora provide evidence of language as it is actually used in specific kinds of communication, by drawing on authentic examples, rather than on assumptions and beliefs: how individuals intuitively think language is used. Corpora are often annotated with grammatical information (i.e. POS—noun, verb, etc.) and/or semantic information (i.e. relating to themes and topics), in addition to metadata regarding where each language excerpt is from (e.g. genre, speaker location). This makes a corpus a valuable electronic tool as it allows users to explore and to better understand a given language.

'National' corpora are typically large in scale (i.e. multi-million or multi-billion words), and often, although not exclusively, contain a combination of written and spoken language (albeit typically a higher proportion of the former to the latter). That is, they are 'general' in their focus, and as they often aim to be 'representative of a particular national language variety' (Hawtin, 2018: 3), they contain texts from a range of different genres, contexts and/or speakers. Given the typical scale and scope (in terms of contents) of national corpora, these datasets are often used as reference corpora, that is, corpora that are used as a baseline against which more specialised (i.e. context or genre specific) and/or small-scale datasets are compared. These comparisons help to ascertain what is similar and/or different to what is 'expected' of a particular language or variety.

The 'first and best-known national corpus' (Xiao, 2008: 384) is the British National Corpus (BNC 1994 – Aston & Burnard, 1997). Built in the 1980s, the BNC 1994 extends to 100-million-words of written and spoken British English, at a distribution of 90 m: 10 m. The BNC 1994 has been used extensively in linguistic research, with a recent search of 'British National Corpus' on Google Scholar reportedly retrieving over 15,800 publications (Love, 2020: 13). A range of other 'national' corpora containing other varieties of English also exists, including those of American, Australian and Pakistani English, as well as 'national' corpora in other European (including Bulgarian, Croatian, Czech, Polish, Slovak, Hungarian, Croatian, Russian and Hellenic) and non-European (e.g. Nepali, Thai) languages. For an extensive list of mega and national corpora, see: http://martinweisser.org, Hawtin, 2018; Love, 2020).

A distinctive feature of 'national' corpora is that they represent language within, crudely speaking, prescribed geographical boundaries and/or political borders. When building a national corpus, decisions may need to be taken about what data within these boundaries/borders to include and exclude and why. The amount of material that might be available to be included in a national corpus will again vary according to the context of the language. Some minority languages such as the French language in Canada or a minoritised language such as Catalan have considerable governmental support and are far better resourced languages than, for example, most of UNESCO's 2500 languages in endangerment (see Moseley, 2010). Planning to build a national corpus would need to reflect the resources likely to be available for its construction and any possible constraints that might arise through this.

Careful consideration will, therefore, need to be given to the sociolinguistic and socio-political context of the language. A language might be used within some specific contexts and/or genres but hardly represented at all in others, for example in academic writing or in the legal system. The available data in some languages might be skewed towards spoken (or written) data and the corpus design would need to reflect this linguistic reality. Similarly, the presence of a language in the e-language mode might be healthy or it might not be very evident. Another important consideration would be the inclusion (or not) of e-language or written data that is predominantly formal and/or translated from a co-official major/majority language. A clear understanding of the contemporary sociolinguistic situation—a 'sociolinguistic audit'—of a language is therefore vital to undertake as the first stage of planning for any national corpus. The following are examples of the types of questions that might comprise part of this sociolinguistic audit, providing a baseline for defining your context:

- How many speakers of the language are there? Are they mainly bilingual or monolingual speakers?
- Where is the language spoken?
- Are other languages spoken within the same territory?
- What is the official status of the language?
- How well-resourced is the language?
- Are there any specific genres that do or do not exist in the language?
- Are there specific contexts in which the language is or is not used in?
- Is the language mainly spoken? Written? Digital?

- Is there a common demographic profile of a speaker of the language? E.g. age, gender, socio-economic background

The answers to these questions will impact on what is feasible to achieve when building a corpus in a given language context.

Language Context: Key Definitions

As already seen, there are many terms for describing languages that are not among the most widely spoken in the world including: under-resourced, minority, minoritised, lesser-used and/or non-dominant languages. It is useful to establish working definitions for these terms here. The term 'under-resourced', in reference to a language, is considered synonymous with the terms 'low density-languages, resource-poor languages, low-data languages, less-resourced languages' (Besacier et al., 2014: 87).

'Resources' might refer to access to basic services, learning materials (for reading, writing, language teaching), pedagogical materials (textbooks and curricula), online opportunities for the language or any combination of these. The absence of these may result in speakers (or users) of that language being at a disadvantage in their educational or cultural experiences. When languages are under-resourced, the population that uses the language exclusively is either somehow linguistically impoverished or forced to acquire an additional language to gain access to the same level of socio-economic status through education or opportunity.

An under-resourced language is 'not the same as a minority language […] some under-resourced languages are actually official languages of their country and spoken by a very large population' (Besacier et al., 2014). A 'minority' language can refer to languages spoken by a minority of the population or whose status is lower than the 'majority' language. A 'majority language', in contrast, can be broadly defined as one that is spoken by most of a population (in a given country or region) and used as the basis for daily communication in most domains, for example, work, education, the media and the law. While under-resourcing is often associated with minority/minoritised languages, it should be noted that for historical and socio-economic reasons, even some widely spoken 'majority' languages might be under-resourced in some contexts. For example,