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

Building a National Corpus

A Welsh Language Case Study

palgrave
macmillan

Contents

1 **Understanding the Language Context** 1
Overview and Aims 2
National Corpora 5
Language Context: Key Definitions 8
Welsh-Language Corpora 14
Contextualising CorCenCC 17
References 19

2 **Design Frames** 23
General Principles 23
Spoken Corpora 28
Spoken Contexts 33
Public/Institutional 36
Media 38
Transactional 40
Professional 41
Pedagogic 42
Socialising and Private 44
Written Corpora 46
Size and Sampling 47
Sampling and Categorising Written Welsh 49
Defining e-Language 55
Genres of e-Language 56
E-Language Corpora 58

Piloting Corpus Building 61
Sampling and Categorising Welsh e-Language 63
Summary 68
References 68

3 (Meta)Data Collection 75
Overviews 75
Participant Recruitment 76
Data Collection 78
Metadata 84
Demographic Targets: Spoken 92
Copyright, Permissions and Ethics 95
References 100

4 Processing and (Re)presenting Corpora 105
Overview 105
Transcribing Spoken Data 106
Automated Anonymisation and Suitability Tagging 114
Tagging 117
Corpus Management Software 120
Y Tiwtiadur 130
References 133

5 Results, Reflections and Lessons Learnt 137
Overview 137
Spoken Data in CorCenCC 138
Reflections and Lessons Learnt I 149
Written Data in CorCenCC 152
Reflections and Lessons Learnt II 158
E-Language Data in CorCenCC 160
Reflections and Lessons Learnt III 164
Concluding Remarks 166
References 168

Appendices 171

Index 181

List of Figures

Fig. 1.1 Key stages of corpus building explored in this book 5
Fig. 2.1 Workflow for creating corpus design frames 68
Fig. 3.1 Promotional coaster used to recruit contributors to the CorCenCC corpus 77
Fig. 4.1 Workflow for transcribing spoken corpora 112
Fig. 4.2 User login options in CorCenCC's database management tool 121
Fig. 4.3 Examples of CorCenCC Data Management Tool interfaces 122
Fig. 4.4 KWIC output for '*dydd*' in CorCenCC (30 random instances from 11,057) 125
Fig. 4.5 Word identifier functionality in CorCenCC's *Y Tiwtiadur* 132
Fig. 5.1 CorCenCC's spoken demographics compared to the 2011 census (ONS, 2011) 147

List of Tables

Table 1.1 Selected under-resourced language corpora and corpus projects 12
Table 2.1 Original broad design frame for CorCenCC 26
Table 2.2 Design frame for spoken 'public/institutional' contexts in CorCenCC 37
Table 2.3 Design frame for spoken 'media' contexts in CorCenCC 39
Table 2.4 Design frame for spoken 'transactional' contexts in CorCenCC 41
Table 2.5 Design frame for spoken 'professional' contexts in CorCenCC 42
Table 2.6 Design frame for spoken 'pedagogic' contexts in CorCenCC 43
Table 2.7 Design frame for spoken 'socialising' contexts in CorCenCC 45
Table 2.8 Design frame for spoken 'private' contexts in CorCenCC 45
Table 2.9 Design frame for books in CorCenCC 50
Table 2.10 Design frame for magazines, newspapers and journals in CorCenCC 54
Table 2.11 Design frame for miscellaneous written data in CorCenCC 55
Table 2.12 Design frame for e-language data in CorCenCC 67
Table 3.1 Welsh speakers in each of local authorities in Wales (ONS, 2011) 93
Table 4.1 Top 20 most frequent words in CorCenCC (across all modes) 127

Table 4.2 Collocates of '*gyfer*' at *n*−/+1, listed according to the frequency of co-occurrence 130
Table 5.1 Composition of CorCenCC's spoken sub-corpus 140
Table 5.2 Genres in the CorCenCC's spoken 'broadcast' data 142
Table 5.3 Composition of CorCenCC's spoken 'educational' data 144
Table 5.4 Breakdown of the 'social' data in CorCenCC 145
Table 5.5 Breakdown of the 'private' data in CorCenCC 146
Table 5.6 Location of spoken data contributors (*n* = 515: 298 contributors did not provide this information) 147
Table 5.7 Comparing CorCenCC's written sub-corpus design frame with the final sub-corpus 153
Table 5.8 CorCenCC's written sub-corpus composition (with revised categorisations) 154
Table 5.9 Composition of the e-language sub-corpus in CorCenCC 161
Table 5.10 Website and blog topics of CorCenCC's e-language sub-corpus 161