Best Practices for Spoken Corpora in Linguistic Research

Best Practices for Spoken Corpora in Linguistic Research

Edited by

Şükriye Ruhi, Michael Haugh,
Thomas Schmidt and Kai Wörner

CAMBRIDGE
SCHOLARS
PUBLISHING

Best Practices for Spoken Corpora in Linguistic Research,
Edited by Şükriye Ruhi, Michael Haugh, Thomas Schmidt and Kai Wörner

This book first published 2014

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-6033-6, ISBN (13): 978-1-4438-6033-8

TABLE OF CONTENTS

Part 2: Discussions on Best Practices in Spoken Corpora

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION: PUTTING PRACTICES IN SPOKEN CORPORA INTO FOCUS

ŞÜKRİYE RUHİ, THOMAS SCHMIDT, KAI WÖRNER AND MICHAEL HAUGH

In an age of linguistic research where scholars are making more and more use of digital spoken corpora, a foreground concern of researchers involved in the creation and sharing of language resources is to attain maximum usability, reliability and "longevity" of the resources for present and future researchers in the language sciences. Assuming that data sharing and sustainability have been secured, there is still the fact that individual characteristics of a spoken corpus strongly impact on the kind of research that can be conducted with the resource. One cannot emphasize enough how decisions taken in the methods and techniques developed and deployed in the field of spoken corpora building and sharing eventually affect what linguists (can) say about language represented in corpora.[1]

Turning a gaze therefore toward the very practices in "creating" spoken corpora and corpora repositories and archives, and enriching the venues where these practices are described and discussed, can unravel potential for moving such practices forward in at least two ways. On the one hand, it can help us to address the shared and differing demands that linguists make of spoken corpora. On the other hand, it can be a step toward the development of commonalities amongst practices in spoken

[1] To offer just one obvious illustration, all spoken corpora need to make decisions on what will be represented as units in their transcription conventions. These in turn will influence, for example, quantification of speaker turns and turn lengths, and any analytic conclusions drawn thereof (see Haugh (this volume), Rehbein, Schalowski, & Wiese (this volume), and Schmidt (this volume), for detailed discussions).

corpora research. In both undertakings, however, it is important not to lose sight of the fact that spoken corpora users and developers have (and will always have) their specific research goals and may "cherish" their own way of doing things, often for very good reasons. Indeed, understanding why spoken corpora researchers do the things they do differently is necessary to trace how spoken corpora researchers view their mission(s) in the creation of human language resources and, at a more practical level, understand how they resolve the requirements and constraints/potentialities of their specific research spaces.

One of the requirements for discussing best practices is arguably documentation of the practices in spoken corpora assembly itself. The last ten or so years evidence a surge in discussing trends and issues in corpus assemblage and processing and in providing researchers with guides to good practices in spoken corpus design (e.g., Adolphs & Knight, 2010; Baude et al., 2006; Thompson, 2005; see also Lüdeling & Kytö, 2008). While guidelines undoubtedly function as springboards for researchers in the actual process of corpus creation, they do not–and cannot be expected to–display what goes on in the actual "assembly line". As pointed out by Schmidt and Wörner (2012, pp. xi-xii), very few research outputs are devoted to documenting the process of corpus construction. In other words, we view spoken corpora construction and sharing as major research endeavors that should also be laid open to academic debate in a manner that is more visible than is currently the case in corpus linguistics. In the present volume, our take on the gap is that it is not only practices in corpus construction but also practices in spoken language processing, archiving and sharing that need to be documented and discussed, as methods and techniques implemented in the two are mutually influential.

The present volume thus aims to function as one such venue by approaching the issue of best practices from two angles: the perspective of corpus assembling and issues of annotation and linguistic analysis, and the perspective of corpus assembling in terms of developing spoken language processing, archiving, and dissemination systems. The volume thus lends voice to scholars who are shouldering responsibilities in spoken corpora creation and sharing by listening to the rationales, decisions, and action plans that underlie their research in spoken corpora. One way to do this is to showcase actual practices for the design, creation and dissemination of spoken corpora in linguistic disciplines like conversation analysis, dialectology, sociolinguistics, pragmatics, and discourse analysis. The volume takes stock of current initiatives so that readers may see for themselves how approaches to speech data processing differ or overlap, and observe where and how a potential for coordination of efforts and

standardisation exists. Placing the rationale and the description of practices at the core of the volume, we believe, will be conducive to further development of common and good practices in spoken corpora research.

The volume grew out of a half-day workshop entitled "Best Practices in Speech Corpora in Linguistic Research", which was organized by the editors on May 21, 2012 at the Language Resources and Evaluation Conference in Istanbul.[2] It comprises a selection of revised and extended versions of papers presented at the workshop, as well as contributions solicited from scholars in the field of spoken corpus research more broadly.

1. "Speech" versus "Spoken" Corpora

In the field of human language resources, and in corpus linguistics in particular, a distinction is made between speech corpora and spoken corpora even though the difference may be blurred and the two types of corpora may blend into each other in practice. It is characteristic of speech corpora to be built for "the detailed study of individual sounds and phonetic features" (Tognini-Bonelli, 2010, p. 25) and speech technology applications, major areas of application being automatic speech recognition, speech synthesis and automatic dialogue systems (Wichmann, 2008, p. 187). Speech corpora:

> do not need to be continuous texts, nor do they need to be collected in natural situations […]. Many are recorded in high-tech laboratories in order to maximise the detail on the recorded version. Speech corpora are thus part of the class of special corpora […], those which for one reason or another are not intended to be representative of the ordinary language in its characteristic use as communication. (Tognini-Bonelli, 2010, p. 25)

In contrast, spoken corpora in the sense we are using the term here are principled collections of electronically available, transcribed and annotated audio and/or video recordings of languages or language varieties that aim to represent the language as used by its speakers in naturally occurring communicative contexts (Anderson, 2010). They thus seek to function as reference points for investigating languages and language varieties (Biber, Conrad, & Reppen, 1998; McEnery & Wilson, 2001).

Spoken language corpora research has been gaining momentum in the last two decades thanks to the efforts of computational linguists, speech technologists, and linguists who work with authentic spoken language

[2] http://www.corpora.uni-hamburg.de/lrec2012/index.html.

resources. However, while research outputs stemming from speech corpora and scholarly work on compiling and analyzing written corpora are abundant and have informed the field of spoken corpora (see, e.g., Arısoy, Can, Parlak, Sak, & Saraçlar, 2009), spoken corpora, as noted in the previous section, have not enjoyed the same attention in research on digital human language resources.

Largely in parallel to the speech technology community, linguists from a variety of fields have intensified their efforts to build or curate larger collections of spoken language data. However, while methods, tools, standards and workflows developed for corpora used in speech technology often serve as a starting point and a source of inspiration for the practices evolving in the linguistic research community, it would be an oversimplification to say that speech technology data and spoken language data in linguistic research are merely two variants of the same category of language resources. Too distinct are the scholarly traditions, the research interests and the institutional circumstances that determine the designs of the respective corpora and the practices chosen to build, use and disseminate the resulting data.

As noted above, the emphasis on compiling or curating spoken corpora mainly from impromptu speech and the goal to have spoken corpora serve the linguistics community makes the research enterprise distinctly different from speech corpora. The volume therefore looks at spoken corpora from a decidedly linguistic perspective. It brings together linguists, tool developers and corpus specialists who develop and work with authentic spoken language corpora, and discusses their different approaches to corpus design, transcription and annotation, metadata management and data dissemination. The discussions in the various chapters of the volume are thus geared toward a better understanding of

- best practices for spoken corpora in conversation analysis,
- dialectology, sociolinguistics, pragmatics and discourse analysis;
- spoken corpora designs and corpus stratification schemes;
- spoken corpora designs for pluricentric languages;
- metadata descriptions of speakers and communications;
- data models, formats and workflows for
- spoken language data in linguistic research and possible routes to
- their standardisation;
- tools for manual transcription and annotation of authentic spoken
- data;
- use of (automatic) methods for tagging and annotating authentic
- spoken data in terms of phonetic, syntactic and discursive

- features;
- corpus management systems
- and dissemination platforms
- for spoken corpora;
- integration of spoken corpora from linguistic research into digital
- infrastructures; and
- legal issues in creating, using and
- publishing spoken corpora for linguistic research

In the following we briefly highlight emerging issues and practices, and provide an overview of the foci of the chapters in the volume.

2. Issues in best practices for spoken corpora

In the past, research on spoken language in linguistics –especially in the fields of conversation analysis, dialectology, sociolinguistics, pragmatics and discourse analysis– has relied on audio recordings collected by the researchers for specific projects. More often than not these data were not available for other researchers. The arrival of publicly available spoken corpora is slowly changing the nature of studies conducted in these fields (see, e.g., Adolphs, 2008; Aijmer, 2002; Andersen, 2000; Ries et al., 2000; Rühleman, 2010; Schauer & Adolphs, 2006). This development has its strengths and weaknesses (see Virtanen, 2009 for a discussion on problems in conducting discourse analysis with corpora). While previous data sources may be characterized as specialized corpora (Tognini-Bonelli, 2010), which embody rich information on interactions in a narrow range of communicative contexts, the designs and metadata of large-scale (and in particular reference) spoken corpora have been criticized for presenting impoverished and static contextual information on interactions (e.g. Rühlemann, 2010; Archer, Culpeper, & Davies, 2008).

Current research on spoken corpora is presenting a number of emerging practices in metadata and annotation (tools) that attempt to remedy this state both in terms of types of corpus construction research and annotation tools and metadata understandings (see, e.g, Adolphs, Knight, & Carter, 2011; Schmidt & Wörner, 2009; Taylor, 2011). The issue of contextualizing spoken interaction in corpora is taken up through discussions on (metadata) annotation in this volume. Discussions on research methodologies in spoken corpora research are also addressing this issue by advocating mixed methods (Bednarek, 2011) or deployment of new computational approaches (see, e.g., Santamaría-García, 2011; Choi & Richards, this volume).

A crucial difference between what we may call traditional spoken corpora and present-day electronic corpora is their design principles. In a discussion on the impact of data on corpus linguistic analysis, Moisl (2009) makes the following critical reminder:

> Data is ontologically different from the world. The world is as it is; data is an interpretation of it for the purpose of scientific study. The weather is not the meteorologist's data – measurements of such things as air temperature are. A text corpus is not the linguist's data – measurements of such things as average sentence length are. (p. 876)

We agree with Moisl that data is an interpretation of the world, and underscore that corpora, too, are interpretations of the world of communication for a number of reasons. The selection of what is to be represented of the huge world of communication involves theoretical and empirical decisions on sampling procedures, each step of which is guided by choice of parameters and size considerations. As underscored by Hunston "issues of corpus design and building take us to the heart of theories of corpus linguistics" (2008, p. 166). In other words, corpus design is theory-driven through and through, and design practices, therefore, crucially impact and delimit what corpora can tell us about the workings of language. Furthermore, the very fact that we "record" communication and then transcribe it means "slicing" data and making several abstractions and moves away from the "natural" world. These procedures too are driven by theoretical and practical decisions and call for more intense investigation and debate of existing practices.

In spite of problems concerning what representativeness and balance mean and how they are to be achieved in corpora (see Leech, 2007), the range of communication features to be represented in a corpus, such as the period of occurrence, genre, speaker biographies, the balance of genre tokens, and principles in collecting recordings are expected to be explicit and well-designed for representing variation in language along four parameters: "time, space, society and situation" (Moreno-Fernández, 2005, p.126). These considerations translate themselves into breadth of coverage in present-day corpora. Discussions on practices in corpus design and stratification are taken up especially in Chapters Eight and Nine in the present volume.

Another feature that distinguishes traditional linguistic corpora from electronic corpora is the practice of including metadata, that is, data that describes the language resource in terms of "contents and context to allow a better and more precise retrieval" (MetaGuide 2003, as cited in Wörner, 2012, p. 400). Accuracy, extensiveness and consistency in metadata make

corpora more explicit for the research questions of the linguistic community. Also the sheer size of the data in current orientations in corpus methodologies, which combine quantitative and qualitative techniques, makes metadata indispensible for corpus designs. While standardisation in spoken corpora metadata practices is desirable (Lehmberg & Wörner, 2008), it remains a challenge, and even though there are formats that are widely used, most are customized to fit the demands of individual projects (Wörner, 2012). Discussions on best practices in metadata emerge in several of the chapters in the volume.

Transcription and further linguistic annotation continue to be at the center of debate in linguistic research that deploys spoken language data (Ochs, 1979; O'Connell & Kowal, 1999). This issue is all the more central for spoken corpora for a number of reasons. Achieving a minimum of interoperability is a prime concern for digital human language resources as researchers in linguistics operate with varying transcription conventions so that these somehow need to be interpretable within another framework. Also at the ground level of conducting research with spoken corpora is the issue of segmenting utterances into units that can be quantified and annotated for linguistic features (e.g., turn and utterance boundaries, the syntactic unit of spoken grammar, discourse segments, etc.). While linguistic research on spoken interaction may continue to debate these issues, spoken corpora creators need to develop, if not what may be called standards and common practices (Schmidt, 2011), at least a set of workable conventions that populations of linguists can agree on. These issues also appear as critical points of discussion in Chapters Five, Thirteen, and Fourteen of the present volume.

Alongside standardisation in transcribing spoken language and ways of improving interoperability, annotation proper continues to hold a prominent place in academic debate on spoken corpora. There are differences even in annotation of what might be called the basics such as POS-tagging and morphological parsing (e.g., Atwell, 2008; Schmid, 2008). In this respect Leech's advice that annotation should be "consensual" and "theory neutral" (1993, p. 275) becomes more of an ideal to be pursued than a reality, especially in the context of conversational management and pragmatic annotation. Moving on to more sophisticated levels of annotation–the syntactic, the discoursal and the pragmatic–we observe major theoretical and empirical challenges. To give one illustration on the empirical side, annotation of pragmatic features have largely remained within the area of speech act annotation, but even here annotation techniques often rely on conventionalized linguistic sequences (see, e.g., Weisser, 2002). Chapters Five and Six in this volume address methodologies and

describe implementations of further annotation at the syntactic, discoursal and pragmatic levels, while Chapter Three proposes to move corpus building pragmatic annotation to the metadata level.

Corpora are also shaping significantly the way researchers think about and study language today (Arppe, Gilquin, Glynn, Martin, & Zeschel, 2010; Gilquin & Gries, 2009) and increasingly more researchers are using these data to address various research questions in theoretical and applied linguistics. Within the context of the changing nature of linguistic research, it is thus becoming even more important to disseminate findings about best practice for building and using spoken corpora to the wider community of researchers. Spoken corpora are labor-intensive and expensive digital resources and the spoken corpora designers are following diverse methodologies even though there are converging linguistic and technical approaches. A pressing concern then in spoken corpora research is the development of corpus construction tools and dissemination platforms that can enable researchers to archive their resources and make them available to others. This volume addresses issues related to the large-scale perspective in spoken corpora research in the chapters in Part Two.

Within this wide range of topics, by showcasing current practices in spoken corpora research in diverse languages and language varieties – namely, Czech, English, French, German, Greek, Polish, Russian and Turkish – the present volume addresses critical issues concerning maintenance, dissemination and resource creation. It thus offers readers a wide perspective on avenues toward best practices in the field.

The book brings into closer contact scholars whose specializations have often remained in relatively different streams of the scientific investigation, that is, scholars whose work fall primarily in conversation analysis, pragmatics and discourse analysis but who are involved in spoken corpus compilation, on the one hand, and scholars who also specialize in linguistics but who have been intensively involved in developing various infrastructures for spoken corpora, on the other hand. As noted in the previous section, this combination of scholars brings into better relief the concerns of data providers and data curators in linguistic research.

3. Overview of the volume

The present volume consists of two parts. A number of themes and pressing issues, all of which are inter-related, can be identified in the various contributions to the volume. Highlighting a common concern, we observe that common ground in annotation, which would enable comparability of data resources, sustainability and dissemination appear in

the foreground in several of the chapters. This not surprising, since comparing data both within and across languages is obviously one key research activity in linguistics. The chapters in Part One approach spoken corpora from the angle of corpus design, spoken data transcription and annotation, as conducted in the fields of corpus linguistics and pragmatics. Part Two also discusses design issues in specific spoken corpora research but address computational issues–data formats, ontologies in transcription, and archiving systems in web platforms and portals. In other words, Part Two focuses more on the "infrastructures" that underlie annotation and repository developments.

In Chapter Two, "Building and maintaining the GeWiss corpus–perspectives on the construction, sustainability and further enrichment of spoken corpora. A showcase", A. Slavcheva and C. Meißner describe the construction of GeWiss, which is a spoken multilingual comparable corpus consisting of academic discourse in three languages (German, English and Polish) and German as a foreign language. They discuss issues related to sustainability and annotation, and matters deriving from the publication of the corpus in different legal frameworks.

In "Constructing general and dialectal corpora for language variation research: Two case studies from Turkish", Ş. Ruhi and E. I. Taş discuss the challenges of constructing comparable corpora for geographical and genre variation for pragmatics research in the context of the Spoken Turkish Corpus and the Spoken Turkish Cypriot Dialect Corpus. The chapter first describes design and metadata practices to achieve comparability for conducting variation research. It then details the major transcription conventions that were developed to address display of geographical variation.

A contribution that also highlights the representation of the pragmatics of spoken language in corpora is the chapter authored by T-S. Pavlidou, C. Kapellidi and E. Karafoti, "The Corpus of Spoken Greek" (CSG). Showcasing the work being conducted in the compilation of the corpus, the authors first describe the design of CSG and argue for the need to represent the fine details of spoken interaction in the transcriptions, as is customary in the conversation analytic approach. They argue that practices in spoken corpora research need to consider the research goals of compiling a corpus in the first place when developing standardisation practices that would maintain comparability of resources to serve wider communities of researchers. As will be observed in the chapters in Part Two, standardisation, interoperability and enhancing dissemination of resources are key issues which researchers working at the "managerial" and computational level of spoken corpus resources, so to speak, are

attempting to improve through various text technological research and development of methods for the corpus compilation process and for corpus dissemination.

Moving on to contributions that focus on linguistic annotation of spoken corpora, I. Rehbein, S. Schalowski and H. Wiese in their chapter entitled "Syntactic annotation of spoken language" discuss the annotation of spoken language syntax in KidKo (the Kiezdeutsch-Korpus), which consists of samples of adolescent language in Berlin-Kreuzberg and Berlin-Hellersdorf. Two crucial problems they tackle are the identification and annotation of speech units and disfluencies. They point to the need to maintain "interoperability and comparability with corpora of written language" but also argue that spoken grammar makes its own demands in annotation. To meet these demands, they argue that annotation of spoken corpora should be data-driven. They therefore work on the surface-level description of language use.

Chapter Six, entitled "Computational analysis of turn-taking patterns in interdisciplinary research meetings", addresses annotation from the user end of spoken corpora research. In this chapter, Choi and Richards describe a discourse-level annotation with R. They discuss the implementation of a tagging system for speaker contributions that may then be submitted to statistical analysis to understand social actions. The authors thereby combine the data analytic techniques of conversation analysis and discourse analysis from the perspective of corpus linguistic statistics and showcase a methodology for analyzing the meeting ground of the micro context and the discursive context in spoken corpora.

The chapters in Part Two address infrastructure issues and tool developments in the annotation, analysis, compilation and dissemination of spoken corpora. The discussion opens with the chapter by P. Skrelin and D. Kocharov, entitled "Russian speech corpora framework for linguistic purposes." The authors discuss work that is being carried out on both Russian speech and spoken corpora developed at the Department of Phonetics, Saint Petersburg State University. They particularly address the phonetic annotation system and tools developed, with a view of enabling linguistic analysis of the sound system of the Russian language and incorporating further annotation carried out on the corpora.

In Chapters Eight and Nine we move into detailed descriptions of two spoken corpora, the first from the Slavic group and the second from Romance languages. In their contribution entitled "Building a data repository of spontaneous spoken Czech," L. Benešová, M. Waclawičová and M. Křen present spoken corpus construction work being carried out in the framework of the Czech National Corpus, where the authors both

showcase a spoken corpus and focus on corpus design and compilation issues. They discuss the challenges of achieving representativeness and balance in a corpus that truly reflects spoken language and the challenges of collating a large-scale corpus at the national level. English is arguably the language that has until recently been represented to a greater extent in terms of its various varieties around the world, notably by the International Corpus of English (ICE) project (Greenbaum & Nelson, 1996; http://ice-corpora.net/ice/index.htm). Corpus compilation work for other languages spoken widely around the globe is therefore a welcome development for linguistic research.

In Chapter Nine, O. Ehmer and C. Martinez describe the stages in the workflow and a number of standards in the creation of the Corpus International Écologique de la Langue Française – CIEL-F, a multimodal corpus of global varieties of spoken French. While descriptions of the procedures followed in the collection of recordings is typical of qualitative work in linguistics, researchers reporting on electronic spoken corpora usually do not touch upon this ground level in detail. Thus Ehmer and Martinez's chapter is unique in this respect. Besides detailing practices and standards followed in transcription, metadata and corpus management, they also discuss how recordings were compiled for CIEL-F.

The subsequent chapters in Part Two focus on long-term archiving of spoken corpora and on issues of interoperability. In Chapter Ten, "Best practices on long-term archiving of spoken language data," P. M. Fischer and A. Witt take up the issues of disseminating and curating corpora. With illustrations from the repository system at the Institut für Deutsche Sprache (Institute for the German Language), the authors claim that long-term availability of spoken corpora depends on "the unconditional employment of standardized methods and formats in all facets of data handling and processing." The authors describe the workflow stages in long-term archiving, from pre-processing of resources technical infrastructure maintenance. As is the case in the work described in the chapter by T. Schmidt, Fischer and Witt emphasize the importance of working in close cooperation with data providers and data curators in tackling issues in the workflow.

Chapter Eleven, co-authored by S. Drude, D. Broeder, P. Wittenburg and H. Sloetjes, also adopts a wide angle perspective on spoken corpora research and present an overview of corpus construction, dissemination and archiving practices developed at The Language Archive (TLA) at the Max Planck Institute for Psycholinguistics. The authors discuss issues of tackling resource and annotation diversity in corpus design; they focus on metadata developments within the framework of IMDI and the component

meta-data infrastructure (CMDI), and dwell upon the tools offered for corpus construction and linguistic analysis at TLA.

Chapter Twelve, by H. Hedeland, T. Lehmberg, T. Schmidt and K. Wörner, demonstrates practices that have emerged in Germany. In their chapter entitled "Multilingual corpora at the Hamburg Centre for Language Corpora", the authors first provide an overview of the development of the Hamburger Zentrum für Sprachkorpora (Hamburg Centre for Language Corpora), which grew out of multilingual, parallel, historical and sociolinguistic corpora compilation and research conducted within the framework of the Research Centre on Multilingualism (SFB 538) at Hamburg University, one of the major outputs of which is EXMARaLDA (a system for constructing and analyzing spoken corpora). The authors then discuss the infrastructure being developed and the technical assistance extended to researchers in conducting corpus research and in integrating the corpora for the wider academic community within this infrastructure.

Representation of spoken language in corpora is a major topic of debate in several of the chapters in Part One. Chapter Thirteen, entitled "The use of ontologies as a tool for aggregating spoken corpora", addresses this issue in the context of the Australian National Corpus (AusNC), which is an aggregated collection of corpora, many of which pre-date the establishment of the AusNC. The authors, S. Musgrave, A. Schalley and M. Haugh, propose tackling the problem of comparability of representations that employ different transcription conventions by discussing the development of ontologies for transcription concepts. They illustrate the development of such an ontology on three sets of spoken data in the AusNC and discuss its implications for the representation of the "theoretical commitments" of the original data contributors.

The last chapter in the volume, authored by T. Schmidt, addresses the question as to whether and to what degree there are commonalities in the various linguistic annotation practices in spoken corpora that have emerged since Bird and Liberman's (2001) annotation graph framework. In his chapter entitled, "(More) common ground for processing speech corpora?" Schmidt offers a bird's eye view of these issues based on comparisons between a wide range of corpora (the Database of Spoken German; Schmidt et al., 2013), the Hamburg Centre for Language Corpora (see Chapter Twelve), and those at the Max Planck Institute in Nijmegen (see also Chapter Eleven), and of the Talkbank/CHILDES/Phonbank family (MacWhinney, 2000; Rose, 2012)). Schmidt also takes up the question of how these annotations are organized and used in corpora. With this chapter we get a bird's eye view of recent accomplishments in

interoperability and the major challenges that lie ahead for spoken corpora to be of even more value to the linguistic community.

When asked to comment on the future of corpus linguistics, Leech (2009) highlights the following challenges:

- the still urgent need to document the large number of languages around the world in electronic corpora
- the need to enrich annotation of spoken corpora, especially in terms of its prosodic, semantic, syntactic, pragmatic and discoursal features
- the need for "breakthroughs" in automated transcription and annotation

In this way, he argues, the field can accomplish "sophisticated methods of analysis" (Leech, 2011, pp.165-166). Writing from the end of spoken corpora creation, annotation and dissemination, while we may have not addressed the third challenge, we believe that the present volume offers a comprehensive representation of research and implementation of practices that address spoken corpora creation, annotation and dissemination.

In Part One of this volume, for instance, such challenges are addressed through the use of (adapted) orthographic representation in corpora that is highlighted in Chapters Eight and Nine or the overall efforts to adhere to "theory neutrality" in linguistic annotation underpinning the work described in many of these chapters. It can also be observed in the references to the use of multiple-checks in annotation and transcription, and differences in the sampling of recordings (i.e., whole or complete) made throughout the chapters in Part One.

The various chapters in Part Two also address such challenges through showcasing various tools that can be deployed for corpus assembly and corpus management. In doing so, however, questions arise around whether there remains a need to develop a standardized set of best practices to avoid inadvertent cases of re-inventing the wheel as it were. One productive way forward to ameliorate this problem is to encourage greater use and provision of open source tools. Yet even with greater emphasis on best practices that are open source, as these chapters collectively suggest, there are nevertheless various specific problems those building repositories and archiving systems face, particularly in relation to the markedly divergent approaches to collecting and representing spoken data. One of the challenges facing those building spoken corpora is that more often than not they are required to play a dual role, both as an environment in which corpus-based research can be carried out, and as a repository that safe-

guards and maintains spoken data in a form that will continue to be accessible to future generations. In that respect, the rapid technological changes we continue to experience in this area offers not only significant promise, but also possible dangers for the unwary.

References

Adolphs, S. (2008). *Corpus and Context. Investigating Pragmatic Functions in Spoken Discourse*. Amsterdam/Philadelphia: John Benjamins.

Adolphs, S. & Knight, D. (2010). Building a spoken corpus: what are the basics? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp.38-52). London/New York: Routledge.

Adolphs, S., Knight, D., & Carter, R. (2011). Capturing context for heterogeneous corpus analysis: Some first steps. *International Journal of Corpus Linguistics 16*(3), 305-324.

Aijmer, K. (2002). *English Discourse Particles: Evidence from a Corpus*. Amsterdam/Philadelphia: John Benjamins.

Andersen, G. (2000). *Pragmatic Markers and Sociolinguistic Variation: A Relevance-Theoretic Approach to the Language of Adolescents*. Amsterdam/Philadelphia: John Benjamins.

—. (2010). How to use corpus linguistics in sociolinguistics. In A. O'Keefe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 547-562). London/New York: Routledge.

Archer, D., Culpeper, J., & Davies, M. (2008). Pragmatic annotation. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, *Volume I* (pp. 613-642). Berlin/New York: Walter de Gruyter.

Arppe, A., Gilquin, G., Glynn, D., Hilpert, M., & Zeschel, A. (2010). Cognitive Corpus Linguistics: five points of debate on current theory and methodology. *Corpora, 5*(1), 1–27.

Arısoy, E., Can, D., Parlak, S., Sak, H., Saraçlar, M. (2009). Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing 17*(5), 874-883. Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5071138.

Atwell, E. (2008). Development of tag sets for part-of-speech tagging. In A. Lüdeling & M. Kytö (eds.), *Corpus Linguistics: An International Handbook*, *Volume I* (pp.501-526). Berlin/New York: Walter de Gruyter.

Baude, O., Blanche-Benveniste, C., Calas, M., Cappeau, P., Cordereireix, P., Goury, L., ... Mondada, L. (2006): *Corpus Oraux – Guide des Bonnes Pratiques*. Paris: Presses Universitaires d'Orléans. Retrieved from http://hal.inria.fr/docs/00/35/77/06/PDF/Corpus_Oraux_guide_des_bonnes_pratiques_2006.pdf.

Bednarek, M. (2011). Approaching the data of pragmatics**.** In W. Bublitz & N. R. Norrick (Eds.), *Foundations of Pragmatics. Handbooks of Pragmatics, Vol. 1* (pp. 537-559). Berlin/Boston: De Gruyter Mouton.

Biber, D., Conrad, S., & Reppen, R., (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Bird, S., & Liberman, M. (2001). A formal framework for linguistic annotation. *Speech Commununication, 33*(1-2), 23-60. doi: 10.1016/S0167-6393(00)00068-6

Bucholtz, M. (2000). The politics of transcription. *Journal of Pragmatics, 32*(10), 1439-1465. doi:10.1016/S0378-2166(99)00094-6

Gilquin, G., & Gries, S. Th. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory, 5*(1), 1-26.

Greenbaum, S., & Nelson, G. (1996). The International Corpus of English (ICE) Project. *World Englishes, 15*(1), 3-15.

Hunston, S., (2008). Collection strategies and design decisions. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 1* (pp. 154-168). Berlin/New York: Walter de Gruyter.

Jenks, C. J. (2011). *Transcribing Talk and Interaction: Issues in the representation of Communication Data*. Amsterdam/Philadelphia: John Benjamins.

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf, & C. Biewer (Eds.), *Corpus Linguistics and the Web* (pp. 133-149). Amsterdam: Rodopi.

—. (2011). Principles and applications of Corpus Linguistics. In V. Viana, S. Zyngier & G. Barnbrook (Eds.), *Perspectives on Corpus Linguistics* (pp. 155-170). Amsterdam/Philadelphia: John Benjamins.

Lehmberg, T, & Wörner, K. (2008). Annotation standards. In A. Lüdeling, & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 1* (pp. 484-501). Berlin/New York: Walter de Gruyter.

Lüdeling, A., & Kytö, M. (Eds.). (2008). *Corpus Linguistics: An International Handbook*, *Vol. I*. Berlin/New York: Walter de Gruyter.

McEnery, T., & Wilson, A. (2001). *Corpus Linguistics. An Introduction*. Edinburgh: Edinburgh University Press.

Moisl, H. (2008). Exploratory multivariate analysis. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 1* (pp.876-899). Berlin/New York: Walter de Gruyter.

Moreno-Fernández, F. 2005. Corpora of spoken Spanish language – The representativeness issue. In Y. Kawaguchi, S. Zaima, T. Takagaki, K. Shibano, M. Usami (Eds.), *Linguistic Informatics – State of the Art and the Future. The First International Conference on Linguistic Informatics* (pp. 120-144). Amsterdam/Philadelphia: John Benjamins.

Ochs, E. (1979). Transcription as theory. In E. Ochs & B. B. Schieffelin (Eds.), *Developmental Pragmatics* (pp. 43-72). New York: Academic Press.

O'Connell, D., & Kowal, S. (2009). Transcription systems for spoken discourse. In D'hondt, S., Östman, J-O., & Verheuren, J. (Eds.), *The Pragmatics of Spoken Interaction* (pp. 240-255). Amsterdam/ Philadelphia: John Benjamins.

Ries, K., Levin, L., Valle, L., Lavie, A., & Waibel, A. (2000). Shallow discourse genre annotation in CallHome Spanish. *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000),* Athens, Greece, 31st May-2nd June 2000. Retrieved from http://www.cs.cmu.edu/~alavie/papers/LREC-2000-Clarity.pdf

Rühlemann, C. (2010). What can a corpus tell us about pragmatics? In A. O'Keeffe, & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 288-301). London/New York: Routledge.

Santamaría-García, C. (2011). Bricolage assembling: CL, CA and DA to explore agreement. *International Journal of Corpus Linguistics 16*(3), 345-370.

Schauer, G. & Adolphs, S. (2006). Expressions of gratitude in corpus and DCT data: Vocabulary, formulaic sequences, and pedagogy. *System 34*, 119-134.

Schmid, H. (2008). Tokenizing and part-of-speech tagging. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*, *Volume I* (pp.527-551). Berlin/New York: Walter de Gruyter.

Schmidt, T. (2011). A TEI-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative 1*, http://jtei.revues.org/142.

Schmidt, T. & Wörner, K. (2009). EXMARALDA – creating, analysing and sharing spoken language corpora for pragmatic research**.** *Pragmatics 19*(4), 565-582.

Schmidt, T. & Wörner, K. (2012). Introduction. In T. Schmidt & K. Wörner (Eds.), *Multilingual Corpora and Multilingual Corpus*

Analysis (pp. xi-2). Amsterdam/Philadelphia: John Benjamins.

Taylor, C. (2011). Negative politeness forms and impoliteness functions in institutional discourse: A corpus-assisted approach. In B. Davies, M. Haugh & A. J. Merrison (Eds.), *Situated Politeness* (209-231). London: Continuum.

Thompson, P. (2005). Spoken Language Corpora. In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice* (pp. 59-70). Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora/ [Accessed 02-01-2014].

Tognini-Bonelli, E. (2010). Theoretical overview of the evolution of corpus linguistics. In A. O'Keefe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 14-27). London/New York: Routledge.

Virtanen, T. (2009). Discourse linguistics meets corpus linguistics: theoretical and methodological issues in the troubled relationship. In A. Renouf & A. Kehoe (Eds.), *Corpus Linguistics: Refinements and Reassessments* (pp. 49-65). Amsterdam/New York: Rodopi.

Weisser, M. (2002). SPAACy – A semi-automated tool for annotating dialogue acts. *International Journal of Corpus Linguistics*, *8*(1), 63-74.

Wichmann, A. (2008). Speech corpora and spoken corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook, Volume 1* (pp. 187-207). Berlin/New York: Walter de Gruyter.

Wörner, K. (2012). Finding the balance between strict defaults and total openness: Collecting and managing metadata for spoken language corpora with the EXMARaLDA Corpus Manager. In T. Schmidt & K. Wörner (Eds.), *Multilingual Corpora and Multilingal Corpus Analysis* (pp. 383-400). Amsterdam/Philadelphia: John Benjamins.

Wynne, M. (Ed.). (2005). *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford: Oxbow Books. Retrieved from http://ahds.ac.uk/linguistic-corpora/

PART 1:

CASE STUDIES ON CORPORA DESIGN, ANNOTATION AND ANALYSIS

CHAPTER TWO

BUILDING AND MAINTAINING THE GEWISS CORPUS: PERSPECTIVES ON THE CONSTRUCTION, SUSTAINABILITY AND FURTHER ENRICHMENT OF SPOKEN CORPORA, A SHOWCASE

ADRIANA SLAVCHEVA
AND CORDULA MEIßNER

1. GeWiss – a comparable corpus of spoken academic language

In March 2013, the GeWiss corpus was released.[1] It is the first freely available comparable corpus of spoken academic language and gives to the scientific community a valuable resource of transcripts aligned to audio-recordings and equipped with comprehensive metadata.[2] Its two access options (i.e., via full texts and concordances) are suitable for both qualitative and quantitative research.

The launch of the corpus was preceded by three and a half years of construction work. Numerous decisions had to be made regarding recording, gathering of metadata, transcription, and documentation. Issues of long-term maintenance and hosting had to be dealt with, so that the

[1] The corpus is accessible via the web interface www.gewiss.uni-leipzig.de. It can be used for research and teaching after free registration.

[2] There is ongoing work in creating corpora of academic discourse, e.g. the *euroWiss* project (cf. http://www1.slm.uni-hamburg.de/de/forschen/projekte/eurowiss.html).