Language Learning & Technology
http://llt.msu.edu/issues/june2014/review3.pdf

June 2014, Volume 18, Number 2
pp. 70–74

REVIEW OF *MULTILINGUAL CORPORA AND MULTILINGUAL CORPUS ANALYSIS*

Multilingual Corpora and Multilingual Corpus Analysis

Thomas Schmidt and Kai Wörner (Eds.)

2012

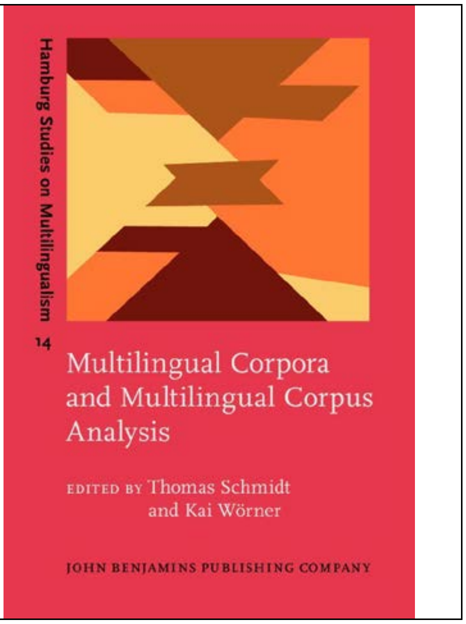
ISBN: 978-9-027-21934-3

US $113.00

407 pp.

John Benjamins

Amsterdam


Hamburg Studies on Multilingualism 14
Multilingual Corpora and Multilingual Corpus Analysis
EDITED BY Thomas Schmidt and Kai Wörner
JOHN BENJAMINS PUBLISHING COMPANY

Review by Nina Vyatkina, University of Kansas

Multilingual Corpora and Multilingual Corpus Analysis is an edited volume that appeared in the Hamburg Studies on Multilingualism series published by John Benjamins. The series focuses on linguistic aspects of multilingualism and primarily publishes studies conducted at the Research Center on Multilingualism at the University of Hamburg, Germany, but also includes outside contributions. This 407-page-long volume edited by Schmidt and Wörner is not an exception in that it comprises 13 studies conducted at the Center and nine studies from external projects. The book is mainly targeted at corpus linguists interested in multilingualism and at learner corpus researchers, which explains the fairly technical language of most studies. However, the volume may also attract attention of multilingualism scholars who are not corpus experts but are looking for empirical data and tools for their research. Furthermore, although several studies report on projects conducted with first language (L1) and/or second language (L2) German corpora, the book also covers a range of other language backgrounds, which expands its target readership.

The book starts with a brief editors' introduction, which is followed by 22 chapters in five sections on 1) learner and attrition corpora; 2) language contact corpora; 3) interpreting corpora; 4) comparable and parallel corpora; and 5) corpus tools. The volume concludes with a general index and also, importantly, a corpora index as well as a language index. Regrettably, the corpora index does not provide internet links for accessing the corpora or relevant proprietary information making it necessary for the reader to search the respective chapters in order to find this information.

The introduction begins with an editors' definition of multilingual corpora that is broader than the one commonly accepted, which mainly refers to parallel corpora with translations of the same texts into different languages. In fact, Schmidt and Wörner's volume includes only one chapter on parallel corpora. The editors expand the definition of a multilingual corpus by referring to "any systematic collection of empirical language data which enables linguists to carry out analyses about multilingual individuals, multilingual societies or multilingual communication" (p. xi). Next, the editors state that their main focus is on methodological issues of corpus design and analysis. Accordingly, the bulk of the volume is devoted to descriptions of available multilingual corpora, presentations of their architecture, and discussions of issues related to corpus design, whereas examples of empirical corpus-based studies are sparse. The editors continue their introduction with a listing of the studies included in each of the five sections. They conclude with a statement that the classification of multilingual corpora reflected in section titles is their own and that the 22 studies comprising the volume could have been arranged differently if other facets of the topic “corpora and multilingualism” were brought to the foreground.

Section 1 is the largest with nine studies and is most directly relevant for language learning research given that it focuses on corpora that comprise L2 data collected from various learner groups in diverse contexts. Studies in sections 2, 3, and 4 may seem more peripheral to language learning research. However, some language learning scholars and language teachers may find the corpora described in these sections useful as a source of linguistic data collected from a variety of multilingual populations. Finally, section 5 describes corpus tools that can be used by scholars who would like to collect or analyze corpora.

Section 1. Learner and Attrition Corpora

The first four studies in this section focus on learner corpora containing data from adult learners in instructed acquisition settings. All these corpora include L2 (or L3) German data, but some of them also contain other L2 data or comparison L1 data. Gut’s opening chapter presents the LeaP (*Lea*rning *P*rosody in a Foreign Language) corpus of spoken learner German and learner English. The LeaP corpus is unique in that it includes spoken samples produced by both native and non-native speakers of two languages in four different speaking styles (free speech, story retelling, and reading-out-loud a story and a list of nonsense words) and is annotated for multiple phonetic, phonological, and prosodic features. More specifically, all speech is graphically represented in a waveform and a spectrogram as well as transcribed orthographically and phonetically. Furthermore, each speech segment is aligned with multiple layers of annotation for: syllable boundaries; consonants, vowels, and pauses; and high and low pitch intervals. Finally, all orthographic words are annotated for parts of speech and lemmas. Gut also shares some research findings enabled by this corpus, namely phonetic correlates of L2 fluency and factors influencing fluency. Hedeland and Schmidt report on methodological issues connected with the creation and annotation of another spoken L2 German corpus (*Ha*mburg *Ma*p *Ta*sk *C*orpus, or HAMATAC) in the second chapter. In particular, they focus on inter-annotator reliability in manual interpretive annotation and create a detailed taxonomy of annotator disagreements. This meticulous analysis shows the importance of reliable annotations for corpus reusability. Ott, Ziai, and Meurers, in the following chapter, also focus on the annotation procedure and inter-annotator agreement as applied to a learner German corpus, but this time to a written *C*orpus of *R*eading Comprehension *E*xercises in *G*erman (CREG). The unique features of CREG are that it contains L2 data at various levels of proficiency, rich task and prompt metadata, and annotations of student responses for comprehension errors based on a compositional semantics taxonomy. The researchers propose ways of improving inter-annotator agreement and call for collecting more task-based corpora that would be valuable for second language acquisition researchers. The next fourth chapter, authored by Zinsmeister and Breckle, focuses on the ALeSKo corpus (*A*nnotiertes *Le*rner*s*prachen*ko*rpus, or Annotated Learner Language Corpus) that comprises L3 German argumentative essays written by advanced learners (with L1 Chinese and L2 English) as well as a comparison subcorpus of L1 German essays. The authors first present the annotation layers of ALeSKo: it is automatically annotated for word classes and lemmatized as well as manually annotated for some discourse phenomena such as local coherence. Next, they report on the results of a comparative study showing that essays written by these fairly proficient learners are still lexically simpler and contain shallower syntactic embedding than native speaker essays.

The next three chapters describe corpora comprised of L2 data produced by young learners. Ulloa, Lleó, and Sánchez describe a collection of L2 Spanish language contact corpora, including spoken data recorded from bilingual children living either in Germany or in Spain, L2 Spanish German children, and monolingual Spanish children. In the next sixth chapter, Lleó adds to this collection by describing two more corpora: one collected from German and Spanish monolingual children and another produced by German-Spanish bilingual children. These corpora are especially valuable because they are longitudinal, spanning data collection periods from two to six years, and are transcribed and annotated for phonetic features, with some annotated for morphosyntactic features as well. Herkenrath and Rehbein address a bilingual Turkish-German and a monolingual Turkish corpus of spoken child language in chapter seven. These three chapters do not report any results of empirical studies, although Herkenrath and Rehbein attempt to present "a methodology for empirical multilingual data analysis" (p. 123). However, due to unconventional terminology use (e.g., "Pragmatic Corpus Analysis"), the value of their chapter for L2 researchers is limited beyond some illustrative snapshots from the corpus.

The final two chapters in Section 1 focus on attrition corpora, defined in Schmidt and Wörner's introduction as "corpora documenting adult speakers' use […] of a first language after a prolonged exposure to a dominant second language" (p. xii). Czachór describes the design of a corpus of Polish data from three groups of L1 Polish speakers: two groups living in Germany, those who had had and those who had never had formal schooling in Poland, and the third being a group of Polish monolinguals living in Poland. The corpus contains a variety of data (spontaneous and elicited speech production, grammaticality judgment, assessment and self-assessment of proficiency, and ethnographic metadata) and is balanced on group size and the amount of data collected per participant. The spoken data is stored as both audio files and transcriptions. Finally, Kupisch, Barton, Bianchi, and Stangen provide an overview of a corpus of audio-recorded and transcribed semi-structured interviews elicited from balanced German-French and German-Italian bilinguals as well as from adult L2 learners. No empirical studies are presented in either chapter on attrition corpora, but interested researchers will find detailed specifications of corpora as well as access information. Notably, all language corpora described in this section are freely available for research with the data owner's consent.

Section 2. Language Contact Corpora

This section focuses on corpora "documenting varieties of languages whose present or historical situation is characterized by language contact" (Schmidt & Wörner, p. xii). It begins with Gabriel's description of a corpus of spoken Argentinian Spanish and a report on some empirical results showing transfer effects from Italian on the speakers' prosody. Next, Kühl presents a corpus containing spoken and written data in Faroese, Danish, and Faroese Danish collected from multilingual speakers of the Faroese Islands. Kühl also reports on a case study showing that language contact effects differ depending on the medium and register (informal spoken versus formal written). In the third chapter, Benet, Cortés, and Lleó describe a corpus of spoken Catalan annotated for auditory and acoustic phonetic features. This corpus contains data from three different age groups (3-5, 19-23, and 32-40 years old) and from two different areas of Barcelona, one characterized by a high degree of contact with Spanish and the other one by a low degree of such contact. The authors also present some research findings showing that the language contact effects were strongest in younger subjects who lived all their life in one district. The fourth chapter by Putz focuses on German dialects of the South Tyrol spoken in the context of contact with Italian, standard German, and Ladin. The production data is comprised of recorded and transcribed medical interactions between dialect-speaking patients and standard-German-speaking physicians, and is manually annotated for instances of misunderstanding occurring due to linguistic factors. The section concludes with Höder's chapter that discusses applications of a syntactically parsed diachronic corpus of Old Swedish to the field of historical linguistics.

Section 3. Interpreting Corpora

This section comprises three chapters describing corpora of recordings and transcriptions of simultaneous or consecutive interpreting. Angermeyer, Meyer, and Schmidt report on the results of a project that involved publishing and sharing community-interpreting corpora in two different settings (doctor-patient communications and court proceedings) and multiple combinations of languages (including English, German, Haitian Creole, Polish, Portuguese, Romanian, Russian, Spanish, and Turkish). The authors do not report on any empirical studies but describe challenges connected with integrating, archiving, and annotating heterogeneous data and propose methods for dealing with these challenges. They conclude with an appeal for sharing community- interpreting data from different projects via a common platform in order to foster interdisciplinary exchange between researchers. House, Meyer, and Schmidt present a smaller specialized corpus of speeches by a Brazilian expert on the topic of genetically modified seeds and their translations into German from Portuguese by different professional interpreters. The corpus contains audio recordings, transcriptions, and annotations for some prosodic phenomena and is freely available. Finally, Bührig, Kliche, Meyer, and Pawlack present a brief description of a corpus of ad-hoc-interpreting in German hospitals. The corpus includes recordings and transcriptions of German-Portuguese and German-Turkish speech samples. The authors also suggest a practical application of the described corpus for training in bilingual workplace communication.

Section 4. Comparable and Parallel Corpora

As the editors explain in their introduction, they include under this subheading corpora that contain texts produced in similar settings but with different languages or language varieties as well as parallel corpora in which original texts are aligned with their translations into other languages. This section begins with the contribution by Fandrych, Meißner, and Slavcheva, who describe a corpus of spoken (monologic and dialogic) L1 academic German, English and Polish, as well as L2 German, currently under construction. This corpus (titled GeWiss) is balanced by genre (academic presentations and oral examinations), sociolinguistic context (native language, second language, and foreign language), and speaker status (academics and students). As the authors point out, this is the first corpus of spoken academic German that can be used for contrastive investigations both with its own sub-corpora and with other academic language corpora such as MICASE (Michigan Corpus of Academic Spoken English) or BASE (British Academic Spoken English). The C4 corpus described in the second chapter by Dittmann, Ďurčo, Geyken, Roth, and Zimmer does not contain L2 data but is a freely and publicly available reference corpus that documents four standard varieties of written German (used in Germany, Switzerland, Austria, and South Tyrol in Italy). It is diachronic as it comprises language samples collected over the course of the 20th century, balanced by decade and text type (journalism, literary texts, scientific literature, and other non-fiction). Furthermore, the corpus is annotated for word classes and is syntactically parsed as well as equipped with integrated search and visualization tools. All these features make the C4 corpus a valuable resource for lexicographic, sociolinguistic, and other linguistic studies as well as German language teaching. Finally, Čulo and Hansen-Schirra describe a parallel corpus that consists of English and German texts from eight different registers with their respective German and English translations. The corpus is annotated for parts of speech, morphological structure, phrase structure, and syntactic dependencies, which makes it suitable for a multitude of linguistic analyses. This chapter is primarily devoted to a technical description of a specific method of syntactic parsing and is oriented toward syntacticians and computational linguists.

Section 5. Corpus tools

The two chapters in this final section do not focus on specific corpora but rather “deal with software tools which can support linguists in compiling and managing multilingual corpora in general” (Schmidt & Wörner, p. xiii). In the first chapter, Rose describes the PhonBank tool that has been developed within the framework of the CHILDES project. This free and open-source tool can facilitate corpus building, oral data transcription and annotation, and phonological analyses of multilingual corpora. In his concluding chapter, Wörner describes the EXMARaLDA corpus building and management tool, specifically focusing on managing metadata (i.e., all data other than primary language data such as information about data sources, data annotations, etc.). This tool is freely and publicly available and allows researchers to add new metadata types to their databases. Notably, most of this volume’s authors have used EXMARaLDA for designing their corpora.

To summarize, *Multilingual corpora and multilingual corpus analysis* is primarily a manual for researchers who wish to design multilingual corpora. It provides detailed descriptions of many corpora that have been collected or are currently being collected in this area. Researchers can choose a model from the many described in this book to suit their specific purposes. However, the book can also serve as a reference book for language learning researchers who wish to use corpora for their data and corpus tools for their analysis. All of the completed corpora described in the book are publicly available on the internet with open access (some of them with the data owner's permission), representing an invaluable resource for researchers. Although many corpora presented here deal with different varieties of German as a first or second language, many other language backgrounds are represented as well. Importantly, all spoken corpora described in the volume (and it predominantly focuses on spoken corpora) include transcriptions and linguistic annotations along with recorded audio data, which makes the spoken data readily available for analysis by external researchers without the need to perform these highly time-consuming activities. Last but not least, corpora described in this book can be used in language teaching as sources of authentic examples or for various language analysis activities that can be designed with corpus tools (see, e.g., Bennett, 2010; McCarthy, 2004; O'Keeffe, McCarthy, & Carter, 2007; Vyatkina, 2013).

ABOUT THE AUTHOR

Nina Vyatkina is an Associate Professor of German Applied Linguistics and coordinator of the German language proficiency sequence at the University of Kansas. Her research interests include longitudinal learner language development, learner corpus analysis, language corpora in language teaching, and interlanguage pragmatics. She is a winner of the 2009 Paul Pimsleur Award for Research in Foreign Language Education (American Council on the Teaching of Foreign Languages / The Modern Language Journal).

E-mail: vyatkina@ku.edu

REFERENCES

Bennett, G. R. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, MI: University of Michigan Press.

McCarthy, M. (2004). *Using a corpus in language teaching*. CALPER Professional Development Document (CPDD-0410). University Park, PA: The Pennsylvania State University, Center for Advanced Language Proficiency Education and Research. Retrieved from http://calper.la.psu.edu/publication.php?page=pdd1

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge, UK: Cambridge University Press.

Vyatkina, N. (2013). Discovery learning and teaching with electronic corpora in an advanced German grammar course. *Die Unterrichtspraxis/Teaching German, 46*(1), 44–61.