

# Contents

Miguel Calderón Campos & Gael Vaamonde

**Introduction. Corpus Linguistics in the Era of Big and Rich Data:  
Methodological Perspectives on Spanish and Portuguese — 1**

## **I Small, Tidy and Rich Diachronic Corpora: The PS-ES and the ODE Corpora**

Gael Vaamonde

**Not so Big Data: Assessing Two Small Specialized Corpora for the Study  
of Historical Variation in Spanish — 11**

Inmaculada González Sopeña

**Language Corpora and Lexical Arabisms in the Digital Age — 37**

Miguel Calderón Campos

**Corpus Size and Tagging: Methodological Strategies for Research on the  
History of Diminutives *-ito*, *-illo*, and *-ico* — 59**

## **II The COSER Corpus and Newspaper Digital Libraries as Alternative Data Sources for Research on Rural and Informal Varieties**

Miriam Bouzouita, Johnatan E. Bonilla & Rosa Lilia Segundo Díaz

**Gaming for Dialects: Creating an Annotated and Parsed Corpus of  
European Spanish Dialects through GWAPs — 87**

María Teresa García-Godoy

**Big Data and Lexical History: Digital Newspaper Libraries in Spanish  
Diachronic Research — 113**

### **III Exploiting Portuguese Reference Corpora: The CdP and the CRPC Corpora**

Amália Mendes

**The Reference Corpus of Contemporary Portuguese: Corpus Design and Case Study on Discourse Markers — 145**

Anton Granvik

**On the Origins of the Shell Noun Construction in Portuguese — 179**

Katharina Gerhalter

***Escrever não escrevo, mas ler um livro, ou um jornal, uns versos, leio. A Corpus-Based Approach to Topicalized Infinitives in Portuguese — 207***