

The Unicode Cookbook for Linguists

Managing writing systems using
orthography profiles

Steven Moran

Michael Cysouw

Translation and Multilingual Natural
Language Processing 10



Translation and Multilingual Natural Language Processing

Editors: Oliver Czulo (Universität Leipzig), Silvia Hansen-Schirra (Johannes Gutenberg-Universität Mainz), Stella Neumann (RWTH Aachen), Reinhard Rapp (Johannes Gutenberg-Universität Mainz)

In this series:

1. Fantinuoli, Claudio & Federico Zanettin (eds.). New directions in corpus-based translation studies.
2. Hansen-Schirra, Silvia & Sambor Gucza (eds.). Eyetracking and Applied Linguistics.
3. Neumann, Stella, Oliver Čulo & Silvia Hansen-Schirra (eds.). Annotation, exploitation and evaluation of parallel corpora: TC3I.
4. Czulo, Oliver & Silvia Hansen-Schirra (eds.). Crossroads between Contrastive Linguistics, Translation Studies and Machine Translation: TC3II.
5. Rehm, Georg, Felix Sasaki, Daniel Stein & Andreas Witt (eds.). Language technologies for a multilingual Europe: TC3III.
6. Menzel, Katrin, Ekaterina Lapshinova-Koltunski & Kerstin Anna Kunz (eds.). New perspectives on cohesion and coherence: Implications for translation.
7. Hansen-Schirra, Silvia, Oliver Czulo & Sascha Hofmann (eds.). Empirical modelling of translation and interpreting.
8. Svoboda, Tomáš, Łucja Biel & Krzysztof Łoboda (eds.). Quality aspects in institutional translation.

The Unicode Cookbook for Linguists

Managing writing systems using
orthography profiles

Steven Moran

Michael Cysouw

Steven Moran , Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles* (Translation and Multilingual Natural Language Processing 10). Berlin: Language Science Press.

This title can be downloaded at:

<http://langsci-press.org/catalog/book/176>

© 2018, Steven Moran , Michael Cysouw

Published under the Creative Commons Attribution 4.0 Licence (CC BY 4.0):

<http://creativecommons.org/licenses/by/4.0/>

ISBN: no digital ISBN

no print ISBNs!

ISSN: 2364-8899

no DOI

Source code available from www.github.com/langsci/176

Collaborative reading: paperhive.org/documents/remote?type=langsci&id=176

Cover and concept of design: Ulrike Harbort

Typesetting: Change typesetter in localmetadata.tex

Proofreading: Change proofreaders in localmetadata.tex

Fonts: Linux Libertine, Libertinus Math, Arimo, DejaVu Sans Mono

Typesetting software: Xe_{La}TeX

Language Science Press

Unter den Linden 6

10099 Berlin, Germany

langsci-press.org

Storage and cataloguing done by FU Berlin

no logo

Preface

This text is meant as a practical guide for linguists and programmers, who work with data in multilingual computational environments.¹ We introduce the basic concepts needed to understand how writing systems and character encodings function, and how they work together.

The intersection of the Unicode Standard and the International Phonetic Alphabet is often met with frustration by users. Nevertheless, the two standards have provided language researchers with a consistent computational architecture needed to process, publish and analyze data from many different languages. We bring to light common, but not always transparent, pitfalls that researchers face when working with Unicode and IPA.

In our daily working lives, we use quantitative methods to compare languages and uncover and clarify their phylogenetic relations. However, the majority of lexical data available from the world's languages is in author- or document-specific orthographies. Having identified and overcome the pitfalls involved in making writing systems and character encodings syntactically and semantically interoperable (to the extent that they can be), we created a suite of open-source Python and R tools to work with languages using profiles that adequately describe their orthographic conventions. Using orthography profiles and these tools allows users to segment text, analyze it, identify errors, and to transform it into different written forms.

We welcome comments and corrections of this text, our source code, and the use case studies that are supplement to this book.² Please use the issue tracker, email us directly, or find the book on PaperHive, where readers can leave questions and comments.³

Steven Moran <bambooforest@gmail.com>

Michael Cysouw <cysouw@mac.com>

¹<https://github.com/unicode-cookbook/cookbook>

²<https://github.com/unicode-cookbook/recipes>

³<https://paperhive.org/>

Acknowledgments

We gratefully acknowledge Robert Forkel, Jeff Good, Jeremy Kahn, Daniel McCloy, Sebastian Nordhoff, and Richard Wright, for inspiration and insights. The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement n° 240816 (PI Michael Cysouw).

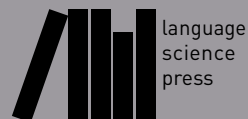
Contents

Preface	i
Acknowledgments	iii
Index	1
Name index	1
Language index	1
Subject index	1

Did you like this book?

This book was brought to you for free

Please help us in providing free access to linguistic research worldwide. Visit <http://www.langsci-press.org/donate> to provide financial support or register as a community proofreader or typesetter at <http://www.langsci-press.org/register>.



Change backtitle in localmetadata.tex

Change backbody in localmetadata.tex

