The Unicode Cookbook for Linguists by Steven Moran and Michael Cysouw

(Language Science Press, 2018. Pp. 132)

Steven Moran is a Professor at the Department of Comparative Linguistics at the University of Zurich. He is trained in language training, linguistic field work, and computational linguistics. Michael Cysouw is a Professor of language typology at Philipps-Universität, Marburg. He is mainly interested in Computational methods and computer-assisted linguistics, Linguistic variation (macro and micro scale), and the processes of language change. Moran and Cysouw introduce this book as a handbook for understanding the way in which writing systems and character encodings function hand in hand by introducing the basic concepts involved in the process of character encodings. The authors mentioned in the preface that the book is for dealing with multilingual computational data and aims to aid linguists and programmers who work together to make languages and linguistic data adequate for computational analysis quantitatively and qualitatively. The writing systems mentioned in the text refer to the orthographical systems of languages. Unicode in its simplest definitions refers to the representation of world's writing systems (Moran & Cysouw, 2018). Whereas character encodings refer to the use of computational tools to interpret, tokenize, and transliterate diverse linguistic sources in a way that makes it meaningful for making comparisons and analysis between languages and various linguistic data. Moran and Cysouw's Unicode Cookbook consists of eight chapters organized in a thematic order. The book addresses experts who already have a sufficient background on how to encode.

The first chapter in its first part introduces how writing systems may differ from one society to another in terms of following the conventional rules of writing. A writing system is a method of representing verbal communication in a visual manner, based on a script and a set of rules regulating its use (Daniels, & Bright, 1996). The authors in the introduction aim to establish that the variability amongst world languages in all levels (orthographically, phonetically, syntactically, etc.) is to be acknowledged and accepted by scholars who must adapt to this variability in order to formalize the orthographical structures for computerizing the writing system in an understandable manner for all. The authors explain that the need for this adaptation arises from language variation. Then the chapter discusses the theoretical background on text encoding, on the Unicode Standard, and on the International Phonetic Alphabet. The notion of encoding is discussed with details in the second part of the first chapter to explain how people exchanged messages in the past by using telegraphy. This reference to the past aims to show how the modern binary encoding system (i.e., a separator between signals and a separator between characters) developed as a consequence for using bisignal codes in the past. The third part of the first chapter discusses the linguistic terminology related to writing systems. In simple words, the authors -in discussing linguistic terminology- aim to make clear distinctions between transcriptions and orthographies to be able to analyze the numerous and various script systems.

Chapter two discusses the Unicode standard to which all digitally encoded operating systems, software, and programming follow. The chapter aims to familiarize the readers to the Unicode approach including the main specification and guidelines for the Unicode standard, the character encoding system, and the representation of grapheme clusters. The chapter starts with a background on the development of the Unicode standard. The background serves as the starting point of the discussion due to its relevance to what comes later in the chapter including the discussion of the different versions of the Unicode standard, the discussion of how characters are represented, the different uses of the properties of the represented characters, the discussion of the main goal of the Unicode standard, and the discussion of the representation of the two categories of grapheme clusters.

Chapter three discusses the Unicode pitfalls in various domains. The chapter aims to highlight where certain technical issues may cause problems. For instance, characters and glyphs are often confused. While a character is the abstract notion of a symbol in the writing systems, a glyph is the visual representation of that symbol (Haralambous & Haralambous, 2003). The problem that might arise from the confusion between a character and a glyph lies in the divergence of one of them because a character could be a piece of a glyph or vice-versa. The chapter discusses the distinctions between various notions that might lead to confusions and technical problems including characters, glyphs, graphemes, rendering (i.e., the substitution of certain Unicode characters), blocks (i.e., the ordering of Unicode characters), names, homoglyphs (i.e., visually indistinguishable glyphs or highly similar glyphs), Canonical equivalence (i.e., the fundamental equivalence between individual Unicode characters and sequences of Unicode characters). Addressing technical issues and problems provides useful information to overcome problems when linguistic data is to be dealt with in computational terms. Indeed, the chapter ends as expected. It ends with recommendations to prevent certain technical problems and issues from emerging, ensure proper linguistic consistency, and deal with various and numerous linguistic diversity (Moran, & Cysouw, 2018, p. 35).

Chapter four aims at spotting the light on the challenges the International Phonetic Association (IPA) faced when the digital encoding of characters first started. The chapter discusses a brief history of the IPA such as how, why, and when it started. The historical overview serves to emphasize that the IPA reflects development in the facts and the theories of phonetic knowledge over time. After the historical view, the chapter discusses the premises and principles of the IPA that developed over time as linguists organized several events to reach common grounds regarding various assignments of characters on the keyboard to the IPA symbols. According to Moran and Cysouw, the IPA created a system that will overcome upcoming challenges including the appearance of new symbols as languages develop.

Chapter five aims to present the encoding issues and propose recommendations for a strict IPA encoding for situations in which cross-resource consistency is crucial. The encoding issues discussed in the chapter include the principles of the IPA in association to creating a single multilingual encoding. In simple words, the issue is how to make characters recognizable by all computers across the globe. The chapter discusses several problems with the IPA principles that have often led to technical encoding issues. The chapter ends in summarizing the pitfalls and recommending defining different types of IPA encodings to force the IPA to commit to a canonical ordering of the characters.

Chapter six aims to guide novice users to the use of special characters in documents. Chapter six is meant to be a practical guide for novice users who are not interested the programmatic aspects discussed earlier. The chapter discusses how certain applications such as Microsoft office and operating systems such as Windows allow special and accented character insertion. Plentiful online resources are provided in the chapter to guide novice users in many aspects of character insertion related to both Unicode and IPA.

Chapter seven aims to guide linguists working with orthography profiles in multilingual environments. The chapter discusses the characterization of writing systems in order to -as the authors state- improve consistency of encoding sources, document knowledge about the writing systems including transliteration, and to do all of that in a way that is quick, simple, and easy to manage for many different sources with many different writing systems. The chapter also discusses the informal description of orthography profiles by suggesting a new proposal that differs from the traditional computational approaches to transliteration. The new proposal -as the authors state- suggests the separation between tokenization (i.e., the task of splitting a stream of characters into words (Habert *et al.*, 1998)) and transliteration (i.e., the process of equating semantic differences with phonetic replacements (Regmi *et al.*, 2010)).

Chapter eight aims to illustrate the practical applications of orthography profiles by introducing the software libraries that provide a practical guide for installing and using them. The chapter discusses issues such as installing the software and overcoming bugs in the system, how to import the library, how to create a tokenizer object, how to tokenize a string, and how to create an orthography profile. The practical tips provided by the chapter simplifies the programming issues handled by programmers on one hand and guides linguists to write profiles and report errors on the other hand.

The book serves it purpose as a handbook for aiding linguists and programmers regarding handling data in multilingual computational environments. Introducing the basic concepts required to understand how writing systems and character encoding function hand in hand is one of the strong features of the book. The links provided in the book regarding certain details the book is not meant to discuss in details are useful. The vocabulary contains technical terms that relate mostly to programmers. However, the book focuses on the Unicode standard but neglects other standards for encoding such as ASCII (American Standard Code for Information Interchange) and UCS-2 (universal Coded Character Set). Even though the Unicode standard is the best we have so far regarding character encoding, yet it is crucial to include other standards of character encoding in order to acknowledge the contribution they have made in the field of computational data. It is also crucial to discuss how these standards handled linguistic data and symbols using variable-width encodings.

References

Daniels, P. T., & Bright, W. (Eds.). (1996). *The world's writing systems*. Oxford University Press on Demand.

Habert, B., Adda, G., Adda-Decker, M., de Marëuil, P. B., Ferrari, S., Ferret, O., ... & Paroubek, P. (1998). Towards tokenization evaluation. In *Proceedings of LREC* (Vol. 98, pp. 427-431).

Haralambous, Y., & Haralambous, T. (2003). *Characters, glyphs and beyond*. Kyoto University 21st Century COE Program.

Moran, S., & Cysouw, M. (2018). *The Unicode cookbook for linguists*. Language Science Press.

Regmi, K., Naidoo, J., & Pilkington, P. (2010). Understanding the processes of translation and transliteration in qualitative research. *International Journal of Qualitative Methods*, *9*(1), 16-26.

HAZIM ALKHRISHEH
University of Pannonia/Hungary
hkhresha@yahoo.com

We acknowledge the financial support of Széchenyi 2020 under the **EFOP-3.6.1-16-2016-00015**.