# Arabic Computational Morphology: Knowledge-Based and Empirical Methods

**Abdelhadi Soudi, Antal van den Bosch, and Günter Neumann (editors)**
(Ecole Nationale de l'Industrie Minérale, Morocco; Tilburg University; and Deutches Forschungszentrum für Künstliche Intelligenz)

Springer (Text, Speech, and Language Technology series, edited by Nancy Ide and Jean Véronis, volume 38), 2007, viii+305 pp; hardbound, ISBN 978-1-4020-6045-8, $159.00

*Reviewed by*
*George Kiraz*
*Beth Mardutho: The Syriac Institute*



The past few decades have witnessed an increased interest in Arabic natural language processing, and in particular computational morphology. In the early 1990s one had to contend with a number of papers that proposed methodologies to handle the various complexities of Arabic morphology, most of which had little implementation associated with them, with the sole notable exception of the works of Beesley, Buckwalter, and others.[1] Today the field has grown, and as this book illustrates, more approaches and implementations are emerging.

The book begins with a preface by Richard Sproat, briefly outlining the early history of Arabic computational morphology. The main body of the book is arranged in four parts. Part I (Chapters 1–3) consists of three introductory chapters. Parts II (Chapters 4–7) and III (Chapters 8–11) present various knowledge-based and empirical approaches, respectively. Finally, Part IV (Chapters 12–15) demonstrates how Arabic morphology is integrated in larger applications, namely, information retrieval (IR), and machine translation (MT). A three-page index lists some grammatical terms and system names.

## 1. Part I: Introduction

The first introductory chapter, by the editors, gives an outline of the book, briefly explaining the various theoretical frameworks on which the rest of the chapters are based. As the book does not contain a chapter on root-and-pattern morphology, the authors discuss this approach in more detail.

The second introductory chapter, by Nizar Habash, Abdelhadi Soudi, and Timothy Buckwalter, introduces the transliteration scheme used in the book (but Chapter 4 does not follow it). The scheme is a one-to-one mapping with Arabic orthographic units (base-line letters and diacritic marks) that is both "complete and easy-to-read" (p. 16). Although Semitic scholars will probably take issue with the "easy-to-read" part of the claim, they can become accustomed to it with ease (by Chapter 6 I was able to read without having to go back to the tables). A new edition of the scheme will benefit from one additional column giving standard Semitic transcriptions, which can be found in any standard work on Semitic grammars.

1 For a review of these early works, see chapter 3 of Kiraz 2001.

The third and final introductory chapter, by Timothy Buckwalter, gives an account of the issues one encounters when dealing with Arabic morphology computationally. The chapter covers orthography (with a detailed account of the history of Arabic encoding systems), orthographic variations, tokenization, lexical design, and dialectical issues.

## 2. Part II: Knowledge-Based Methods

This part of the book presents four knowledge-based methods for handling morphology: syllable-based, inheritance-based, lexeme-based, and stem-based.

The syllable-based approach by Lynne Cahill ("A syllable-based account of Arabic morphology") describes the Arabic tri-literal verbal system using the DATR formalism (`www.datr.com`). The authors are confident that their system will scale to bi- and quadriliteral roots, as well as to weak roots, but do not give any examples in DATR to boost the reader's confidence. An advantage of this approach is that it does not require any additional mechanisms to the existing syllable-based approach.

The inheritance-based approach by Salah R. Al-Najem ("Inheritance-based approach to Arabic verbal root-and-pattern morphology") demonstrates that the Arabic verbal system (both tri- and quadriliteral forms) exhibits a number of generalizations, dependencies, and syncretisms. He demonstrates how these three features can be implemented in DATR. The generalizations are implemented through direct inheritance by placing forms that are more general in higher nodes in the network hierarchy. Dependencies are implemented in a similar fashion where dependents are placed lower in the hierarchy and may require multiple inheritance rules. Syncretisms are implemented by DATR inference rules. As with the previous chapter, the approach does not require any mechanisms additional to what already exists. In both cases, it remains to be seen whether a complete system can be covered in DATR, what the computational and space complexities would be, and how they would differ from the complexities of earlier finite-state approaches.

The lexeme-based approach by Violetta Cavalli-Sforza and Abdelhadi Soudi ("A trade-off between multiple operations and multiple stems") implements Arabic verbal and nominal forms (including the problematic issue of the broken plural) using the Lisp-based MORPHĒ system. The system is driven by a morphological form hierarchy that "describes the relationship of all morphological forms to each other," and transformational rules that attach to leaf nodes in the hierarchy.

The final chapter of this part, "Grammar–lexis relations in the computational morphology of Arabic," by Joseph Dichy and Ali Farghaly, is more challenging to read and seems disjointed. Without reading the abstract, the reader will wonder what the objective of the chapter is until the very end. In the words of the volume editors, the chapter "provides an in-depth discussion of the role of grammar–lexis relations . . . After presenting the limits of [previous systems], the authors argue that entries associated with a finite set of morphosyntactic w-specifiers can guarantee a complete coverage of the data within the boundaries of the word-form" (p. 8).

## 3. Part III: Empirical Methods

The first chapter, "Learning to identify Semitic roots," by Ezra Daya, Dan Roth, and Shuly Wintner, embarks on resolving a difficult task, the recognition of roots from

surface forms using a statistical machine-learning approach. The authors begin with a Hebrew system and extend it to handle Arabic. They report a precision of over 80%, which they compare to the average human performance for the same task.

The second chapter, "Automatic processing of modern standard Arabic text" by Mona Diab, Kadri Hacioglu, and Daniel Jurafsky, demonstrates how Arabic texts can be processed in terms of tokenization, lemmatization, part-of-speech tagging, and base phrase chunking. To achieve this, the authors employ a support-vector-machine learning approach, and extend its traditional use to tokenization. The accuracy results they report range from 91.6% for base-phrase chunking to 99.1% for clitic tokenization.

The third chapter, "Supervised and unsupervised learning of Arabic morphology," by Alexander Clark, experiments with learning the Arabic broken plural using a general-purpose learning algorithm. The algorithm makes use of non-deterministic stochastic finite-state transducers that perform transductions between two surface forms, inflected and non-inflected.

The final chapter in this part, "Memory-based morphological analysis and part-of-speech tagging of Arabic," by Antal van den Bosch, Erwin Marsi, and Abdelhadi Soudi, describes a memory-based learning technique for morphological analysis and part-of-speech tagging. They report a joint accuracy for both tasks of 58.1%. They conclude that although memory-based approaches are feasible for morphological analysis, they are unable to recognize the stems of unknown words. They note that the approach, however, works well for part-of-speech tagging.

### 4. Part IV: Integration in Larger Applications

The last part of the book consists of two chapters on IR and two chapters on MT.

The first chapter, "Light stemming for Arabic information retrieval," by Leah S. Larkey, Lisa Ballesteros, and Margaret E. Connell, is a preliminary study on light stemming, where inflectional variants are conflated into one stem. They argue that light stemming is adequate for IR purposes. Their stemmer simply removes common prefixes and suffixes from words after they are normalized by removing their diacritics.

The second chapter, "Adapting morphology for Arabic information retrieval," by Kareem Darwish and Douglas W. Oard, also describes stemming for IR. Here, the authors use an existing morphological analyzer to build root-word pairs. In the training phrase, they align segments of the root with their counterparts in the surface form. The result is used statistically to recognize the stems of new words. A light stemmer that removes prefixes and suffixes is also developed, and both modules were tested on LDC data.

The third chapter, "Arabic morphological representations for machine translation," by Nizar Habash, outlines the issues one faces when dealing with Arabic MT. Habash focuses on the morphological representation of Arabic in statistical and rule-based MT systems. He evaluates a bidirectional system called Almorgeana where a feature-set (that includes a stem) is used to generate one or more Arabic words, and words are analyzed to give the same feature-set. Habash tests this system on a sample text of over one million words in diacritized and undiacritized modes.

Finally, the last chapter of the book, "Arabic morphological generation and its impact on the quality of machine translation to Arabic," by Ahmed Guessoum and Rached Zantout, concentrates on morphological generation of Arabic words for MT. The chapter enumerates a number of common types of errors to be found in Arabic MT systems, and provides an evaluation of a commercial MT system against this list.

## 5. Summary

In conclusion, this collection of essays is essential for any researcher interested in Arabic morphology and demonstrates how far the field has grown since the early and mid 1990s. Throughout the book, the reader will no doubt come across numerous rule-based and statistical NLP approaches that may not be familiar. The authors made sure that these approaches are well explained in each chapter, with sufficient references should the reader wish to dig in further.

In a collective work on a language whose morphological data is quite complex and by far differs from "the norm" we find in Western languages, authors are required to give a description of the data at hand, and they *all* did. The result is a repetitive description of the same data, with some variations depending on the topic of each chapter. This could have been avoided by collapsing the description in an introductory chapter, and having the authors reference their work to this common description. This would also have forced authors to relate their descriptions and results to each other, giving a more unified presentation. This, however, is a very minor point and should not take away from the excellence and richness of the individual presentations.

## References

Kiraz, George. 2001, *Computational Nonlinear Morphology, With Emphasis on Semitic Languages*. Cambridge University Press, Cambridge, UK.

*George Kiraz* is the founder and director of Beth Mardutho: The Syriac Institute and the president of Gorgias Press. He earned an MSt in Syriac Studies from Oxford University, and an MPhil and PhD from Cambridge University. He has an extensive list of publications in Syriac studies and is the author of *Computational Nonlinear Morphology, With Emphasis on Semitic Languages* (Cambridge University Press, 2001). Kiraz's address is Gorgias Press LLC, 180 Centennial Ave., Suite A, Piscataway, NJ 08854, USA; e-mail: gkiraz@gorgiaspress.com.