

Reviews

Linguistic Identity Matching. By BERTRAND LISBACH and VICTORIA MEYER. Wiesbaden, Germany: Springer Verlag, 2013. Pp. xvii + 244. \$59.99. ISBN 978-3-8348-1370-1.

Every day, around the globe, millions of people perform at least one electronic identity search for a service, an organization, a group, or an individual. Despite the frequency of this action, few of us have ever really taken the time to consider the staggeringly complex system of computer programming required by the rapid search and accurate retrieval of identity data — that is, until something goes wrong, for example, when our search produces far too few useful matches (true positives) or far too many irrelevant hits (false positives). When performing private, small-scale, informal electronic identity search-and-retrieval operations, such as trying to find the name of a long-lost high-school sweetheart, problems in hit rate and search accuracy can be extremely annoying. However, in the high-stakes professional worlds of governmental record-keeping, global finances, and international law enforcement, the failure to perform global name matches accurately and reliably, quickly and efficiently, can have devastating, even life-threatening, consequences. The book under review provides readers with a fascinating look into some of the long-term challenges and recent solutions for name-matching processes for computerized identity matching.

With this goal in mind, the book is organized into three conceptual parts. In Part I, readers are introduced to the basic concepts, terminology, and application of linguistic identity matching. Here, for example, readers are introduced to the differences between “natural” and “legal” persons, as well as distinctions among “database profiles,” “search profiles,” and “person searches.” After these computerized identity matching (CIM) concepts and real-world applications are presented, Part I goes on to examine the ways in which socio-historical, orthographic, and morphophonological variations in personal names may affect CIM processing. To illustrate these effects, the authors present a series of concrete examples from a range of languages and cultures. For onomastic scholars within academia, this comparative presentation may make large segments of this work quite useful for teaching graduate courses in computer linguistics, anthropological linguistics, and of course onomastics. For onomastic scholars outside of academia, however, what will be of interest here is not so much the book’s description of this cross-linguistic variation but rather the technical challenges this onomastic diversity presents.

To address these challenges, the second part of the book goes on to describe the leading name-matching (NM) methods that have been used historically. Described briefly, these methods are divided into three generations. This segment therefore introduces readers to industry standards such as Levenshtein distance, N-grams pattern matching, and generative algorithms. Using a series of onomastic examples taken again from a variety of language families, the authors methodically compare and contrast the various generations of NM methods and thereby effectively illustrate the successive improvements that have been made over the years. This historical examination offers a solid foundation for the authors’ conclusion that the key to maximizing the true potential of CIM lies in elevating not only the technical but also the linguistic techniques applied in NM. As the authors note in their introduction:

Until now, the majority of the literature [...] has focused on identity matching as a purely technical use and often failed to highlight the linguistic aspects that are vital for any effective solution. This book is intended to introduce the successful use of computation linguistics for identity matching to a wider audience and to encourage the attention that this relatively new science undoubtedly deserves. (xvii)

Both together and apart, the information and examples presented in Parts I and II offer a convincing argumentation for this underlying supposition.

The third and final part of the book then goes on to demonstrate the multiple ways in which modern NM methods are combined with other non-linguistic features or “attributes” to increase the overall recall and precision of CIM. The authors even offer valuable tips for evaluating the efficacy of CIM software and provide a well-argued rationale for implementing linguistic search standards for CIM enterprises. Importantly, this discussion will be of interest not only to onomasticians with expertise in anthroponymy. Researchers who specialize, for example, in toponymy may also be interested in reading about the ways in which toponymic variation can affect the precision of CIM operations. For onomasticians with little or no experience with the technical aspects of database design, programing, and management, there is no reason to fear. This reference provides a logical conceptual foundation that succeeds in educating without boring or overwhelming the reader. Therein lies one of the truly outstanding features of this work. One need not have an advanced degree in computer science or information science to understand and enjoy the material presented.

Of course, no reference is without its shortcomings, and this reference is by no means an exception to the rule. For example, in segments involving highly theoretical issues, it is regrettable that sample case studies are not used. It is naturally understood that employing authentic individual-level name data would have been impossible given the privacy and security issues involved. However, mock case studies based on real-life NM and CIM projects would have significantly increased the accessibility of some of the more technical, less immediately accessible, concepts. This pedagogical technique would have been far more effective than the lengthy summaries provided at the end of each chapter.

For the diehard technophobe, there are admittedly certain sections of this work which may still remain stubbornly inscrutable. Another shortcoming of this reference is the tendency of the authors to rely repeatedly on the exact same onomastic data to illustrate different methodological challenges. For example, the Russian name *Yeltsin* is used multiple times to discuss a number of different issues, e.g. transcription variation (10), cross-linguistic deviation in gender-marking (37), the influence of data-entry specialists’ native languages (54), similarity and edit operations in Levenshtein distance calculations (93), N-grams (98), Soundex methods (95), and thesaurus-based matching methods (102). Obviously, such repetition can become rather annoying, especially when reading more than a few chapters at a single sitting. That having been said, from a pedagogical perspective, one might also say that this repetition could be extremely beneficial for increasing readers’ comprehension and retention of important concepts. It could also be argued that using the same examples to illustrate different CIM problems effectively underscores the necessity for developing programs powerful enough to address the multifaceted onomastic data simultaneously.

Whichever position one decides to take, it generally can be agreed that this rhetorical idiosyncrasy does little to mar the overall scholarly value of this reference. In the final analysis, this publication not only makes a substantive contribution to applied onomastics, a subject area which still receives far too little attention today, but also calls important attention to a societal issue that has recently gained profound international significance, namely the hidden advantages and disadvantages of widespread CIM methods for the individual and society. Precisely this point is alluded to by David Smith, the forensic accountant and CIM specialist who writes the foreword to this reference. As Smith keenly observes, at the time that this book was written, the revolutions and uprisings in the Middle East had produced large-scale sanctions necessitating high-level, sophisticated CIM. At that time, experts recognized, the concurrent dangers of producing incorrect matches while overlooking correct ones were exponentially magnified by the real-life sociopolitical stakes involved. What was true then may be even more so today. For that reason, it is essential that more scholars with onomastic expertise become aware of and proficient in the ways that names can be and are being used for CIM purposes. In the final analysis, for onomasticians, the true value of this reference is not so much in the linguistic insights it provides but in the education it offers regarding the increasing global demand for high-speed, low-error automated CIM methods and

applications. Although this reference certainly will not provide all of the information needed to understand CIM technology and practices completely, it is an excellent starting point. *Linguistic Identity Matching* provides a fascinating look into a world that most of us come into contact with each and every day without ever even noticing it.

Germanic Society for Forensic Linguistics (GSFL)

I. M. NICK



A Dictionary of First Names. By PATRICK HANKS, FLAVIA HODGES, and KATE HARDCASTLE. 2nd edn. Oxford: Oxford University Press. 2006. Reissued in Oxford Reference Collection, 2016. Pp. xxviii + 434. \$35.00 (hardback). ISBN 978-0-19-880051-4 © American Name Society 2017.

Onomasticians and linguists know Patrick Hanks as the former chief editor of current English dictionaries at Oxford University Press (1990–2000) and editor of the *Oxford Dictionary of Family Names in Britain and Ireland*, with Richard Coates and Peter McClure (2017), and the *Oxford Dictionary of American Family Names* (2003). With Flavia Hodges, he published the *Oxford Dictionary of First Names* in 1990. A second edition, in collaboration with Kate Hardcastle, appeared in 2006 and now has been reissued in the Oxford Reference Collection. The new format makes this valuable compendium accessible handily (~5¼ × 8 in.) and economically, in keeping with the announced goal of the series: “The Oxford Reference Collection uses sustainable print-on-demand technology to make the acclaimed backlist of the Oxford Reference Programme perennially available in hardback format” (pre-title page).

The heart of *A Dictionary of First Names* remains the entries of over 6,000 first names, defined as “the first of a sequence of one or more given names borne by an individual,” with the corollary explanation that a “given name is one that is bestowed on a child by its parents or guardians at birth, as opposed to an inherited surname” (xi). The distinction between a first given name and a second or other given name becomes important as a criterion for inclusion in the work. As the authors explain, second and third and additional names are drawn from a much wider range, often including mothers’ surnames or choices unique to the family of origin. Only once such a name has achieved enough popularity to serve as a common given name, such as *Douglas* or *Dudley*, is it included in the dictionary (xi).

Entries in *A Dictionary of First Names*, printed two columns per page, interfile male and female names, identified with the symbol ♂ and/or ♀ immediately after the name, in continuous alphabetical listings. The name’s description follows the gender marker, typically including the language from which the name is derived, its meaning, and, as available, such information as historical point of origin, periods of popularity, and famous historical or literary figures who carry the name. When warranted, alternate forms and spellings are listed at the end of the entry, i.e. short form, pet form, and variant(s). Entries do not include pronunciation.

Two medium-length entries are given below as representative. Note that they embed cross-references, indicated by an arrow and change in font, that direct readers to related entries within the work.

Kelsey ♀, ♂ Transferred use of the surname, which is from an Old English masculine personal name *Cēolsige*, derived from *cēol* ‘ship’ + *sige* ‘victory’. Its use as a girl’s name may have been influenced by such names as ►Elsie. In the United States the spelling *Kelsey* is reserved chiefly for boys.

VARIANTS:Kelsi(e) ♀. (156)

Sandra ♀ Short form of *Alessandra*, the Italianate form of ►Alexandra. A major influence in establishing this as a common given name in the English-speaking world was George