Book Review

Machine-Aided Linguistic Discovery: An Introduction and Some Examples

Vladimir Pericliev
(Bulgarian Academy of Sciences)

London: Equinox, 2010, ix+330 pp; hardbound, ISBN 978-1-84553-660-2, $90.00, £60.00

Reviewed by
Eric J. M. Smith
University of Toronto

The subtitle of Vladimir Pericliev's book, *An Introduction and Some Examples*, is a succinct and accurate description of its contents. Pericliev argues briefly for the usefulness of computer-aided techniques in linguistic discovery, contrasting it with the intuitionist approach which has characterized linguistic discovery throughout much of its history. The bulk of the book is devoted to examples of software-aided linguistic discovery drawn from his own work.

Chapter 1 starts by sketching out the current state of discovery techniques in linguistic theory, categorizing scientific discovery into three main approaches: the intuitionist approach, the chance approach, and the problem-solving approach. Discoveries by intuition and by chance remain the purview of humans, but clearly the problem-solving approach can benefit from the application of computational techniques.

Chapter 2 presents the KINSHIP program, which performs "parsimonious discrimination" in order to determine the minimal set of features which are necessary to discriminate all of a language's kinship terms. The program is used to discover feature geometries, superior to existing human-discovered ones, which describe the kinship terminology of languages like English and Bulgarian.

Chapter 3 extends the ideas used in KINSHIP to a program called MPD (maximal parsimonious discrimination), which is then applied to a variety of other tasks, some of which are unconnected to linguistics. Of these applications, the most interesting is the use of MPD to determine the segment profiles which uniquely identify languages in the UPSID-451 database (consisting of segment inventories from 451 languages, selected to provide broad coverage of the world's language families) (Maddieson and Precoda 1991). Although Pericliev discusses his results at considerable length, it is not clear what the theoretical usefulness of these profiles might be. What does it really tell us about French to know that it is the only language in the database to contain the phoneme [œ̃]? Of more practical interest was Pericliev's discussion of the process of converting the UPSID data into a featural representation to make it amenable to processing, describing how to represent underspecified segments and how to deal with transcription variations. This sort of necessary preprocessing constitutes an important and underemphasized part of the process of machine-aided linguistic discovery. The study of UPSID does produce some interesting, though not unexpected, results. For instance, when a profile contains more than one unique segment, the majority of these segments share a common feature, and 85.8% of the unique segments have some sort of secondary articulation.

Chapters 4 and 5 present two more pieces of software developed by Pericliev: UNIV and AUTO. The UNIV software is inspired by Greenberg's universals (Greenberg 1966), and automates the haphazard process by which universals have been identified in the past. Given a vector of features for each language being studied, UNIV identifies all universal patterns which hold above a user-specified threshold. Such universals can be unrestricted or statistical and they can be stand-alone or implicational. Once a set of universals has been identified, the results are fed through AUTO (for "AUthoring TOol"), which assembles boilerplate text into a journal article; given the vast number of (often trivial) universals which UNIV discovers, this can be useful. UNIV is first applied to two data sets: one of kinship terms, and the other the word-order data used by Greenberg himself. The most interesting result is that Greenberg's set of word-order universals was neither complete nor fully supported by the data.

UNIV is then applied to the UPSID-451 database. UNIV identifies a large number of previously unnoted universals, most of which are rather low-level and of little inherent theoretical interest. However, the low-level machine-discovered generalizations can then be used as the basis for more interesting manually created generalizations. The UNIV analysis also serves to refine earlier claims made by Maddieson (1984) and Gamkrelidze (1978).

Chapter 6 is devoted to MINTYP, which is a program for determining the minimum typology to account for an observed set of universals. The search for such typologies is discussed by Greenberg (1966) and by Hawkins (1983). MINTYP takes a system of universals and a set of logically admissible types, and eliminates any superfluous universals which can be implied by stronger universals in order to determine the smallest set of universals which still accounts for the observed data. This approach is able to distill Greenberg's set of universals into as few as four composite universals. Like UNIV and KINSHIP, MINTYP follows Pericliev's basic approach: Reduce the data to a set of features, and then find the patterns which most economically cover the observed feature distribution.

Chapter 7 turns this featural approach to the problem of genetic language classification with the RECLASS software. Pericliev extends the featural approach to include Swadesh-type word-lists, for which he describes a method for calculating a similarity metric based on phonological features of words in the list. A set of languages from different families is selected for study, a similarity metric is calculated for pairs of languages, and unrelated languages whose similarity is significantly greater than expected are given further attention. In Pericliev's test case, the initial feature data consists of kinship terminology, which revealed an unexpected similarity between the Kaingang languages of Brazil and various Polynesian languages. He pursues this similarity first by using features based on word-list similarities and then by looking at other structural features, arguing at length for the plausibility of a genetic connection. Of all the results described in the book, this is probably the most interesting, because it represents a discovery made by Pericliev's machine-aided approach which is unlikely ever to have been found by the haphazard manual process of discovery.

The main weakness of the book is that all the software described in the book was developed or co-developed by Pericliev himself, so the various programs all risk seeming like variations on a single theme. The book would have benefited by including examples of software from others working in the field, which might differ from the feature-coverage approach favored by Pericliev. That being said, Pericliev's essential point is a valid one: Machine-aided discovery has a tremendous untapped potential for analyzing data sets which are too large to be amenable to human inspection. The success of this approach is best exemplified by his machine-aided discovery of a possible genetic relationship which would otherwise have eluded human discovery.

References

Gamkrelidze, T. V. 1978. On the correlation of stops and fricatives in a phonological system. In J. H. Greenberg, editor, *Universals of Human Language*. Stanford University Press, Stanford, CA, pages 2:9–46.

Greenberg, J. H. 1966. Some universals of grammar with particular reference to the order of meaningful elements. In J. H. Greenberg, editor, *Universals of Language*. Mouton & Co., The Hague, pages 73–113.

Hawkins, J. 1983. *Word Order Universals*. Academic Press, New York.

Maddieson, Ian. 1984. *Patterns of Sounds*. Cambridge University Press, Cambridge, UK.

Maddieson, I. and K. Precoda. 1991. Updating UPSID. *UCLA Working Papers in Phonetics*, 74:104–114.

Eric Smith's primary research at the University of Toronto is centered on corpus approaches to the study of Sumerian syntax. Earlier machine-aided discoveries included the reconstruction of Elamite phonology using Optimality Theory. His e-mail address is `eric.smith@utoronto.ca`.