Title

Review of Measuring Linguistic Differences

Author

Operstein, Natalie

Publication Date

2024-06-30

Peer reviewed

REVIEWER: Natalie Operstein
AUTHOR: Borin, Lars & Saxena, Anju (eds)
TITLE: Approaches to Measuring Linguistic Differences
SERIES TITLE: Trends in Linguistics Studies and Monographs, Vol. 265
PUBLISHER: De Gruyter Mouton
YEAR: 2013

SUMMARY

“Approaches to Measuring Linguistic Differences”, edited by Lars Borin and Anju Saxena, contains selected contributions to the Workshop on Comparing Approaches to Measuring Linguistic Differences that was held at the University of Gothenburg in October 2011. It consists of an introduction and nineteen chapters arranged in two sections, "Case Studies" and "Methods and Tools". In the following summary, the order of presentation of the individual chapters differs from the order in which they appear in the volume.

In the introductory chapter, "The why and how of measuring linguistic differences", Lars Borin introduces the volume by surveying the field of linguistic distance measurement and summarizing the individual volume contributions. The focus of the survey is on the choice of features on which computation of linguistic distances is based (e.g. lexical versus grammatical) and on the methodology of distance calculation.

Two of the chapters address the measuring of linguistic distance between the source and target languages in the context of language learning. In "Predicting language-learning difficulty", Michael Cysouw explores ways of quantifying the difficulty of a particular language from an English speaker's point of view. The measured distances between English (as L1) and various target L2s include geographical, genealogical, orthographic and structural (the last one based on the data from WALS). Some of the results confirm preexisting (presumably, empirical) data on the comparative difficulty of foreign languages for English speakers, specifically the Foreign Service Institute's language difficulty ranking (http://www.effectivelanguagelearning.com/language-guide/language-difficulty). "The effect of linguistic distance across Indo-European mother tongues on learning Dutch as a second language", by Job Schepens et al., complements the perspective of Cysouw's study by focusing on a single target L2 (Dutch) and multiple source languages. The authors find that the difficulty of learning Dutch as a second language is positively correlated with the linguistic distance between Dutch and the learners' L1s.

Two chapters address the use of vocabulary lists for computationally supported classification of languages. In "Carving Tibeto-Kanauri by its joints: using basic vocabulary lists for genetic grouping of languages", Anju Saxena and Lars Borin use a revised, 157-item Swadesh list for investigating the internal relationships of a subgroup of the Tibeto-Burman language family. Their study also provides the context for an interesting general discussion pertaining to the compiling and use of basic vocabulary lists for the genetic classification of languages. In "Using semantically restricted word-

lists to investigate relationships among Athapaskan languages", Conor Snoek uses a list of 61 body part and fluid terms to investigate relationships among 23 Athapaskan languages and dialects. The distance between the languages is calculated based on the orthographic forms of the relevant items in the word lists. Among the interesting findings of the study is that of particular diachronic stability of the terms for body fluids.

Two chapters compare the results of internal classification of language families obtained by manual and computational methods. In "How aberrant are divergent Indo-European subgroups?", Folke Josephson discusses the findings of Ringe et al. (2002) against the background of current discussion of subgrouping relationships in the field of Indo-European linguistics. In "Distance-based phylogenetic inference algorithms in the subgrouping of Dravidian languages", Taraka Rama and Sudheer Kolachina apply quantitative methods to unresolved issues in the internal classification of Dravidian. As part of their study, they create four datasets -- three lexical and one structural -- for a large number of Dravidian languages.

Two chapters quantitatively assess differences between varieties of the same language. In "Degrees of semantic control in measuring aggregated lexical distances", Kris Heylen and Tom Ruette study lexical variation in a corpus of Dutch-language texts published between 1999 and 2004 and composed of four subcorpora: Usenet posts, popular newspaper articles, quality newspaper articles, and government announcements. The specific focus of their study is on calculating the lexical distance between the subcorpora by using a set of 218 variables represented by 476 variants (e.g. the variable "car" is represented by the variants 'auto' and 'wagen'). In "Measuring socially motivated pronunciation differences", John Nerbonne et al. focus on elucidating the place of regiolects (regional speech variants) in the Dutch dialect-standard continuum. The authors use a modified version of the Levenshtein distance (Levenshtein 1965) to measure pronunciation differences between the base (local) dialects, standard languages (Belgian and Netherlandic Dutch) and regiolects, as exemplified by the speech of eight regional radio announcers employed by radio stations in the Netherlands and Flanders. They find that of the eight regiolects examined, only one occupies an intermediate position between the base dialects and the standard. Of the remaining seven, five were found to be more different from the base dialects than the standard, and two were more different from the standard than the base dialects. These results lead the authors to conclude that the primary function of regional speech is not so much to facilitate communication as to express social identification with the region.

Two chapters focus on comparing the performance of different computational methodologies when applied to the same dataset. In "Black box approaches to genealogical classification and their shortcomings", Jelena Prokić and Steven Moran evaluate three fully automatized approaches to genetic language classification. Their dataset consists of 366-item word lists in 69 indigenous languages of Colombia, classified into 12 language families. The approaches in question do not require any linguistic knowledge beyond the division of words into segments, and are based on the Levenshtein distance (computed as the smallest number of character insertions, deletions and substitutions required to transform one word into the other in their pairwise

comparison), the n-gram analysis (which computes the similarity between two words as the number of shared fixed-length strings divided by the length of the word), and the zipping approach (which estimates the distance between texts in two languages by merging the texts and measuring the compression rates). The authors find that, although all the techniques were able to capture higher-level genetic groupings (with the zipping approach being less accurate than the other two), none of the methods is able to capture deeper genetic relationships or provide other information useful to a historical linguist, such as probable cognates or recurrent sound changes or sound correspondences. Another study aiming to compare two methods of measuring linguistic distance -- cognate counting and the Levenshtein distance -- unexpectedly produced unanticipated results. In "Languages with longer words have more lexical change", Søren Wichmann and Eric W. Holman report their finding that the rate of lexical replacement is positively correlated with the length of words, both within and across languages. The authors' provisional proposed explanation appeals to cross-linguistic differences in word-formation techniques.

Several chapters attempt to quantify semantic distance. In "Word similarity, cognation, and translational equivalence" Grzegorz Kondrak begins with the observation that "words that are phonetically similar across different languages are more likely to be mutual translations" (p.375). This observation finds empirical support in two case studies involving similar French-English word pairs, one based on a bilingual dictionary and the other on a bitext. The author hypothesizes that the similarity of mutual translations is explainable, in part, by their similar frequency in the respective languages, which translates itself into similarity in word length.

In "Measuring morphosemantic language distance in parallel texts", Bernhard Wälchli and Ruprecht von Waldenfels base their approach on the assumption that "forms with most similar distributions across parallel texts also have similar meanings" (p.475). E.g., they find that the English pronoun '*he*' occurs in 2319 verses of the New Testament while the corresponding German pronoun '*er*' occurs in 2014 verses, and attribute this considerable distributional overlap to "the highly similar use of the two forms across the two languages" (p.476). The backbone of the approach is pairwise comparison of parallel texts in a large number of languages; the case studies reported in the paper involve a world-wide sample of 60 parallel texts from the Bible and a more restricted sample of 27 translations of the novel '*Master and Margarita*' (plus the Russian original). The paper is rounded off with a discussion of potential applications of the proposed language distance measure, most notably to typological sampling.

In "Semantic typologies by means of network analysis of bilingual dictionaries", Ineta Sejane and Steffen Eger endeavor to give numeric expression to lexical semantic distance between languages. Their method is based on using bilingual dictionaries in multiple languages to generate semantic network representations of a common reference language. The influence of the reference language is held to be invariant, and the differences in the segmentation of lexical semantic space are attributed to the languages used for creating the networks.

In "Comparing linguistic systems of categorisation", William B. McGregor proposes and tests possible ways of measuring cross-linguistic distance with respect to a grammatical construction. The construction in question, found in some languages of northern Australia, is a type of complex predicate consisting of a coverb (uninflecting verb) and an inflecting verb. After introducing the construction and surveying how it has been approached in the literature, McGregor considers three possible distance measures that abstract away from the generic properties of the construction and center instead on its less predictable, more individual lexical and semantic properties: the relative size of the coverb sets that collocate with a particular inflecting verb, the degree of similarity of the collocating coverb sets, and shared versus language-specific coverb / inflecting verb collocations (pp.398-399).

Several chapters focus on presenting or testing the possibilities of specific research tools, resources or methodologies. In "Towards automated language classification: a clustering approach", Armin Buch et al. apply the software tool CLANS (CLuster ANalysis of Sequences) to several case studies to illustrate the use of clustering approaches to genetic language classification. One of the case studies involves 46 parallel translations of the Bible in 37 languages and endeavors to measure the degree of syntactic similarity between languages. The latter is quantified, e.g., via a version of the Levenshtein distance which measures the extent of difference in the linear order of mutual translations.

In "Information-theoretic modeling of etymological sound change", Hannes Wettig et al. propose a methodology for automated extraction of recurrent sound correspondences from pre-compiled cognate sets. The methodology is tested by using two digital Uralic etymological resources as the input. The paper also provides a context for a general discussion of shortcomings and possible biases inherent in the manual compilation of etymological datasets.

In "Contrasting linguistics and archaeology in the matrix model: GIS and cluster analysis of the Arawakan languages", Gerd Carling et al. outline an interdisciplinary project utilizing GIS (Geographic Information System) technology to collate spatially distributed linguistic and non-linguistic data. The focus of the study is on relationships among the Arawak languages as well as Arawak cultural patterns. The linguistic data consists of both lexical and structural features, while the non-linguistic component includes data from archaeology, ethnohistory, ethnography and physical geography.

In "Dependency-sensitive typological distance", Harald Hammarström and Loretta O'Connor aim to quantify typological distance between languages in a way that incorporates dependency relationships among the features. They test the proposed distance metrics on a database consisting of 81 binary features, 42 morphosyntactic and 39 phonological, in 35 Chibchan and neighboring languages from the Isthmo-Colombian Area.

In "The Intercontinental Dictionary Series -- a rich and principled database for language comparison", Lars Borin et al. describe the origin and structure of the

Intercontinental Dictionary Series (IDS), an electronic lexical database freely available online at http://ids.clld.org/. The database is composed of 1,310-item word lists, with 215 such lists available at the time of the publication. The entries fall into three categories -- universal concepts, cultural concepts and geographical / environmental phenomena -- and are identified using English words for the respective word senses, with the list of the word senses modeled after Buck (1949). The authors illustrate the use of IDS lists for linguistic research by referencing a project that investigates the Himalayan region as a possible linguistic area, partly with the help of this tool (https://spraakbanken.gu.se/eng/research/digital-areal-linguistics).

EVALUATION

The volume presents an interesting collection of articles endeavoring to give numeric expression to the elusive notion of the amount of distance between languages. The articles exemplify a number of linguistic subfields where such an endeavor might be useful, ranging from the practical (language learning and teaching) to the theoretical (typological sampling, genetic and typological classification of languages, subgrouping of language families). Many additional potential applications are pointed out in the individual contributions (e.g. p.500). The volume's chapters discuss, explain and/or compare a number of different methodologies, tools and resources and present creative proposals on how to use different aspects of the language, including phonology, syntax, semantics and the lexicon, to convert the observed cross-linguistic and/or intra-linguistic differences into tangible distance metrics.

A useful feature of several of the chapters are abundant references to past and current literature on the techniques of measuring linguistic distances. Engagement with concrete data also provides the framework for addressing larger conceptual and methodological issues in linguistics. These include, but are not limited to, the causes of language change, genetic stability of typological features, grouping of languages by their segmentation of semantic space, potential biases in the construction of etymological datasets, variable understanding of the notion of 'basic vocabulary' in different subfields of linguistics, improvement of lexically-based dating techniques, methodological problems in the compilation and use of vocabulary lists, and how fully automatized analyses of data compare with manual analyses of the same data as performed by experts. The overall effect of this discussion is to contribute to the distillation of techniques and data types potentially amenable for use as linguistic distance markers.

The volume's focus on linguistic differences also contributes to the growing body of literature on linguistic complexity (e.g. Miestamo et al. 2008; Sampson et al. 2009), while at the same time serving as a timely reminder about how different the world's languages really are. To quote Wälchli and von Waldenfels, "[l]anguages are more different from one another than the average linguist believes" (p.492). The volume presents a variety of numerical approaches to measuring different aspects of these differences for various purposes ranging from language learning to linguistic typology and to language classification.

REFERENCES

Buck, Carl Darling. 1949. A Dictionary of Selected Synonyms in the Principal Indo-European Languages: A Contribution to the History of Ideas. Chicago: University of Chicago Press.

Levenshtein, Vladimir. 1965. Dvoichnye kody s ispravleniem vypadenij, vstavok i zameshchenij simvolov [Binary codes with correction of deletions, insertions and symbol replacements]. Doklady Akademii Nauk SSSR 163: 845-848.

Miestamo, Matti, Kaius Sinnemäki and Fred Karlsson (eds). 2008. Language Complexity: Typology, Contact, Change. Amsterdam/Philadelphia: John Benjamins.

Ringe, Don, Tandy Warnow and Ann Taylor. 2002. Indo-European and computational cladistics. Transactions of the Philological Society 100: 59-129.

Sampson, Geoffrey, David Gil and Peter Trudgill (eds). 2009. Language Complexity as an Evolving Variable. Oxford: Oxford University Press.