An entropy-based assessment of the UNICODE encoding for Tibetan

Paul G. Hackett

This paper presents an analysis of the UNICODE encoding scheme for Tibetan from the standpoint of morphological entropy. We can speak of two levels of entropy in Tibetan: syllable-level entropy (a measure of the probability of the sequential occurrence of syllables), and letter-level entropy (a measure of the probability of the sequential occurrence of letters). Syllable-level entropy is a purely statistical calculation that is a function of the domain of the literature sampled, while letter-level entropy is relatively domain independent. Letter-level entropy can be calculated statistically, though a theoretical upper bound can also be postulated based on language dependent morphology rules. This paper presents both theoretical and statistical estimates of letter-level entropy for Tibetan, and explores the Tibetan UNICODE encoding scheme in relation to coding ambiguity, data compression, and other issues analyzed in light of an entropy-based language model.