# Decomposability and Frequency in the Hindi/Urdu Number System

**Chundra Aroor Cathcart (chundra.cathcart@ling.lu.se)**
Department of Linguistics and Phonetics
Lund University

**Abstract**

Hindi/Urdu (HU) numbers 10–99 are highly irregular, unlike the transparent systems of most languages. I investigate the morphological decomposability of HU numbers using a series of computational models. While these models classify most forms accurately, problems are encountered in high-frequency forms of low cardinality, suggesting that some HU numbers are more transparent (i.e., morphologically decomposable) than others. These results are compatible with a dual-route access model proposed for the processing of numeral forms.

**Keywords:** numerals; computational modeling; Bayesian learning; Hindi/Urdu; phonology; morphology

## Introduction

Hindi/Urdu (HU, officially considered separate languages but differing in little other than orthography and high-register vocabulary) and most other modern Indic languages are unusual in that the numbers through 99 are highly opaque and irregular, undoubtedly posing difficulties in production and processing for language users. This phenomenon is understudied, and raises interesting questions regarding the design principles of cross-linguistic number systems, as well as the processing of complex morphological forms by language users.

In this paper, I investigate the mechanisms by which HU users extract structure and meaning from HU number terms. I address this issue using a series of unsupervised computational models designed to approximate HU users' processing of the numerals 10–99. While it is generally agreed that number terms are acquired as individual lexical items, there is good reason to hypothesize that many numbers above a certain threshold of frequency are not accessed as individual lexemes during processing, but rather via their component parts, in line with a dual-route model of lexical storage and access (cf. Baayen, 1993). I expect that despite the irregularity of the HU number system, users can find regular patterns in various cues to numerical identity in the input, particularly in low-frequency numbers, thus facilitating easier comprehension.

My methodology investigates the extent to which HU numbers can be morphologically decomposed. Brysbaert (2005) tentatively proposes a dual-route access model for English numeral storage, hypothesizing that frequent, opaque items like *twelve* are accessed directly, while morphologically transparent numbers of lower frequency (e.g., *eighty-nine*) are processed through decomposition. In line with this view, I predict that less frequent HU numbers can be segmented and labeled more accurately by a computational model, indicating greater morphological transparency.

I find that a model using *n*-grams as phonological features successfully assigns most HU numeral forms to the proper TENS/DIGITS cohort, but that, rather unsurprisingly, some highly opaque forms are misclassified. Major errors occur among numbers of lower magnitude. Since these forms are highly frequent, this state of affairs is compatible with a dual-route account of processing. I find that in general, the model faces difficulties in capturing relationships between simplex (i.e., monomorphemic) forms (e.g., /əssi/ '80') and their complex counterparts (e.g., /cɔrɑsi/ '84'), where a more sophisticated model of phonology might succeed. These results provide an important baseline for future investigations into mental representations of HU numerals.

## Background

The full list of numerals (taken from Comrie, n.d.) is given in Table 1. When encountering a datum like /bəjɑlis/ '42', listeners must infer the value of the TENS and DIGITS place with the aid of cues in the input, and must be able to contend with highly noisy allomorphy: TENS$\{40\}$ and DIGITS$\{2\}$ have multiple surface realizations. In some cases, this allomorphy is suppletive (i.e., variants bear no phonological resemblance to each other). Listeners may possess the knowledge that HU is head-final, and that higher-order numerical information generally occurs closer to the root (Hurford, 1987), i.e., to the right. For some numerals, it seems plausible that high frequency facilitates access; for instance, HU /sola/ '16' is quite unlike other numerals with the feature DIGITS$\{6\}$, all of which are /c$^{\text{h}}$/-initial. This is a diachronic artifact; /sola/ faithfully continues Sanskrit *ṣoḍaśa-*, while other forms with DIGITS$\{6\}$ contain reflexes of an unattested dialectal variant *kṣ(v)aṭ- of attested Sanskrit *ṣaṣ-* '6' (Turner, 1962–1966). It is also the only member of the teens which shows /l/ in its allomorph of /dəs/ 'ten'. All the same, it may be used frequently enough that this twofold suppletion does not pose problems to speakers and listeners.

A major attempt to explore synchronic regularities among HU numbers is that of Bright (1969), who concludes that despite a lack of economy, implicit rules governing the system are available to language users. Berger (1992) outlines the complex historical development of HU numbers; sporadic phonological reduction, analogy, and language contact, among other phenomena, have resulted in a highly irregular and opaque system compared to the relatively transparent numbers of Sanskrit, HU's ancestor. These works aside, many aspects of the HU numeral system remain untreated.

## Representational issues

### Abstract representation of HU numerals

Above, I adopt the canonical abstract numerical representation found in much of the literature, where each surface form comprises two underlying factors corresponding to the TENS and DIGITS place. I make the assumption that DIGITS$\{0\}$

Table 1: HU numbers 1–99; rows represent the tens place, columns the digits place

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | — | ek | do | tin | car | põc | $c^h\varepsilon$ | sɑt | $ɑt^h$ | nɔ |
| 10 | dəs | gjɑrə | bɑrə | terə | cɔdə | pəndrə | solə | sətrə | $ət^hɑrə$ | ʊnnis |
| 20 | bis | ɪkkis | bɑis | teis | cɔbis | pəccis | $c^həbbis$ | sətːɑis | ətːɑis | ʊntis |
| 30 | tis | ɪkəttis | bəttis | tɛ̃tis | cɔ̃tis | pɛ̃tis | $c^həttis$ | sɛ̃tis | ərtis | ʊntɑlis |
| 40 | cɑlis | ɪktɑlis | bəjɑlis | tɛ̃tɑlis | cəvɑlis | pɛ̃tɑlis | $c^hɪjɑlis$ | sɛ̃tɑlis | ərtɑlis | ʊncɑs |
| 50 | pəcɑs | ɪkjɑvən | bɑvən | tɪrpən | cəuvən | pəcpən | $c^həppən$ | səttɑvən | $ətːt^hɑvən$ | ʊnsə$t^h$ |
| 60 | sɑ$t^h$ | ɪksə$t^h$ | bɑsə$t^h$ | tɪrsə$t^h$ | cɔ̃sə$t^h$ | pɛ̃sə$t^h$ | $c^hɪjɑsə t^h$ | sərsə$t^h$ | ətsə$t^h$ | ʊnhəttər |
| 70 | səttər | ɪkhəttər | bəhəttər | tɪhəttər | cəhəttər | pəchəttər | $c^hɪhəttər$ | səthəttər | $ət^hhətər$ | ʊnjɑsi |
| 80 | əssi | ɪkjɑsi | bəjɑsi | tɪrɑsi | cərɑsi | pəcɑsi | $c^hɪjɑsi$ | səttɑsi | $ətːt^hɑsi$ | nəvɑsi |
| 90 | nəve | ɪkjɑnve | bɑnve | tɪrɑnve | cərɑnve | pəcɑnve | $c^hɪjɑnve$ | səttɑnve | $ətːt^hɑnve$ | nɪmjɑnve |

does not map to any overt phonological information. Additionally, for forms such as /ʊntɑlis/, there is a mismatch between the abstract representation TENS{3} DIGITS{9} and the phonological form, since the morpheme representing the tens place closely resembles /cɑlis/ '40', not /tis/ '30'; this suggests an intermediate calculation TENS{4} DIGITS{−1}. I assume that the representation DIGITS{−1} is an integral part of HU numerical computation and is reflected explicitly in the morphology.

## Surface representation of HU numerals

Brysbaert hesitates to draw a categorical distinction between transparent and opaque English numerals, citing semi-transparent forms like *thirteen*. Along these lines, I seek to situate HU numerals along a cline between mild and extreme opacity. I quantify a number's transparency or decomposability via the performance of a computational model designed to segment and label HU numbers, both in terms of (1) accuracy of the labeling and (2) low posterior uncertainty.

At the outset, I lack a principled means of separating suppletive and non-suppletive allomorphy found in the system. Numbers 11–18 exhibit three allomorphs for TENS{1}, /-də/, /-rə/ and /-lə/, all from the diachronic source *-daśa-*, though synchronic /d/ ∼ /r/ ∼ /l/ alternations are not well known in HU. Numbers 49–58 show multiple bases for TENS{5}, all descended from Sanskrit *pañcāśat-* '50' but formally very dissimilar. I make no a priori assumptions about the status of suppletive allomorphy in the morphological system, and allow the model to simply group together forms according to the configuration it infers. I do, however, treat DIGITS{−1} as a separate morpheme, given its systematic occurrence. Nonparametric models (which assume an unbounded number of underlying morphological labels) may alleviate some of the problems that result from forcing suppletive allomorphs to be classified together, which I set aside for future work.

A model of HU numerical processing should characterize the morphological structure of the data encountered. HU numbers are highly fusional, exhibiting the effects of millennia of phonological and morphological change. As Bright reports, no economical set of rules helps to derive the surface representations from their morphological bases. There is often unpredictable allomorphy between simplex and complex forms of a given decade: for instance, /s/ alternates with /h/ in complex forms of /səttər/ '70' < Sanskrit *saptatí-*, but not in complex forms of /sa$t^h$/ '60' < Sanskrit *ṣaṣṭi-*; however, the latter decade's complex forms contain a reduced vowel /ə/, alternating with /ɑ/. The short vowel and geminate consonant found in /əssi/ '80' alternate with a long vowel and singleton consonant in derived /-ɑsi/, but the short vowel found in /nəve/ does not appear in derived forms.

Despite these challenges, listeners should be able to form a probability distribution over possible morphemes contained in a complex input datum. HU's highly fusional phonology notwithstanding, listeners should be able to approximate the location of morpheme boundaries. This question is a key part of this paper's computational inference, and should be of broad interest to phonological theory, as it has the potential to incorporate a number of strategies for morphological boundary detection. The models introduced in this paper draw morpheme boundaries on the basis of what is most likely under the current parameters of the model, and are dependent on distributional information found in other numerals. This task is easier than that of many types of unsupervised segmentation in that at most one boundary must be located per input datum; however, the model must contend with a wider distribution of allomorphs which must be unified. This model does not use external distributional information for the purpose of segmentation (as do Harris, 1955; Saffran et al., 1996).

A question relevant to this paper concerns the types of morphological segmentation that should be permitted. Cross-linguistically, a morphological segmentation of the type [b][əjɑlis] might be permissible, but non-inflectional morphemes in HU tend to consist minimally of a unit with prosodic weight. As such, I restrict the proposal distribution for segmentations of HU numbers to exclude morpheme boundaries following the first and penultimate segments; this additionally speeds up inference and ensures that short forms like /bis/ '20' will be treated as monomorphemic.

## Phonological features

The methodology developed in this paper must capture allomorphy in the HU numeral system, inferring that different surface strings such as /-jɑlis/ and /cɑlis/ correspond to the same underlying morpheme, TENS{4}. I cluster allomorphs together on the basis of phonological features using essentially the same likelihood formula used in unsuper-

vised Naïve Bayes/Dirichlet-Multinomial classifiers, popular in bag-of-words models of document classification. This is a model of convenience which I find fairly effective, though it is admittedly crude; it is insensitive to positional information and alternations, and does not strongly penalize the absence of potentially crucial morpheme-level information.

The model described in this paper lends itself to the use of different types of phonological features, and provides opportunities to investigate model performance under features with differing degrees of abstraction. In this paper, I limit myself to domain-general string-based features, namely *n*-grams. In many contexts, I expect segmental bigrams to fare well in assigning cohort membership. However, there are some cases where I fear that bigrams may fail to capture alternations caused by (among other things) deletion or insertion, as between the base /nəʊe/ '90' and /-nʊe/, found in complex forms. A unigram model will be sensitive to the co-occurrence of /n/ and /ʊ/, whereas a representation consisting solely of bigrams will not (on the use of separate autosegmental tiers for consonants and vowels, see Goldsmith & Riggle, 2012). I attempt to circumvent this problem with a phonological representation that uses both unigrams and bigrams (though this technically violates the independence assumption of Naïve Bayes). This allows the model to capture some similarities between paradigmatically related forms that would otherwise be lost in a strict bigram model.

## Model

Here, I introduce the core model employed in this paper, designed to approximate a HU speaker's recognition of numbers 10–99 (I assume that 1–9 are primitives). When encountering a numerical form, the listener must determine whether it is simplex or complex. If simplex, the value of the TENS place must be inferred; if complex, the DIGITS place must be as well. The model assumes that a complex form is generated by independent draws from two mixtures, a DIGITS mixture (the labels of which correspond to the values $\{-1,1,...,9\}$) and a TENS mixture (the labels of which correspond to the values $\{1,...,9\}$). Because HU morphology is generally concatenative, I make the simplifying assumption that phonological elements generated by a given mixture are adjacent to one another — i.e., that a morpheme boundary can be located somewhere in a complex form, however approximately. I make the assumption that the lefthand morpheme is generated by the digits mixture and the righthand morpheme is generated by the tens mixture; this convention essentially incorporates Hurford's insight that higher numerical elements occur closer to the root, which in turn can be interpreted as prior knowledge of a morphosyntactic headedness parameter. This system of numerical classification is schematized in Figure 1.

### Inference

This paper's basic model of numeral classification assigns each form to one or two mixtures, given a $10 \times F$ matrix $\Omega^D$ and a $9 \times F$ (where $F$ is the number of feature types in the input) matrix $\Omega^T$ (specifying a prior over feature distributions associated with each label of the DIGITS and TENS place, respectively), as well as a word-level vector $\mu$ representing a prior over morpheme boundary locations. I initialize these matrices with symmetric concentration parameters $\alpha^T \alpha^D, \alpha^\mu$, set to .1 in order to encourage sparseness, such that unshared features from unrelated labels are not clumped together. The generative model draws probability simplices $\phi_j^D \sim \text{Dirichlet}(\omega_j^D), \phi_i^T \sim \text{Dirichlet}(\omega_i^T)$ representing the feature distributions associated with levels $j$ and $i$ of the DIGITS and TENS place, and assumes that for every word $w$,

$\varsigma \sim \text{Dirichlet}(\mu)$ (a simplex of morpheme boundary probabilities is drawn, including the probability $p(m = \emptyset)$, i.e., the probability that there is no morpheme boundary)
$m \sim \text{Categorical}(\varsigma)$ (a morpheme boundary is drawn from $\varsigma$)
If $m = \emptyset$,
  $z_j^D = 0$
  for each feature $f \in w$
    $f \sim \text{Categorical}(\phi_i^T), i \in \{1,...,9\}$
If $m \neq \emptyset$,
  For each feature $f \in w_{1,...,m}$ (through index $m$)
    $f \sim \text{Categorical}(\phi_j^D), j \in \{-1,1,...,9\}$
  For each feature $f \in w_{m+1,...,|w|}$ (from index $m+1$ through the end of the word)
    $f \sim \text{Categorical}(\phi_i^T), i \in \{1,...,9\}$

I marginalize out the parameters $\varsigma, \phi_i^T, \phi_j^D$ to obtain collapsed Dirichlet-Categorical updates for $p(m|\mu), p(z^T|\Omega^T), p(z^D|\Omega^D)$. For a given word, this yields the following conditional probability if $m = \emptyset$ (adopted from Yin & Wang, 2014):

$$P(m = \emptyset, z_i^T, z^D = 0 | z_{-i}^T, z_{-0}^D, \Omega^T, \Omega^D, \mu) \propto \frac{\prod_{f \in w} \prod_{n=1}^{c(f)_w} c(f)_{z_j^T}^{-w} + \alpha^T + n - 1}{\prod_{k=1}^{|w|} c(\cdot)_{z_j^T}^{-w} + F\alpha^T + k - 1} \quad (1)$$

If $m \neq \emptyset$:

$$P(m, z_i^T, z_j^D | z_{-i}^T, z_{-j}^D, \Omega^T, \Omega^D, \mu) \propto \frac{\prod_{f \in \lambda_m^l} \prod_{n=1}^{c(f)_{\lambda_m^l}} c(f)_{z_j^D}^{-w} + \alpha^D + n - 1}{\prod_{k=1}^{m} c(\cdot)_{z_j^D}^{-w} + F\alpha^D + k - 1} \cdot \frac{\prod_{f \in \lambda_m^r} \prod_{n=1}^{c(f)_{\lambda_m^r}} c(f)_{z_i^T}^{-w} + \alpha^T + n - 1}{\prod_{k=1}^{|w|-m} c(\cdot)_{z_i^T}^{-w} + F\alpha^T + k - 1} \quad (2)$$

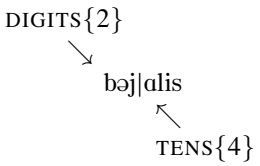


Figure 1: Schema of a proposed morphological segmentation, tens classification, and digits classification for form /bəjɑlis/

Above, $c(f)_{z_i^T}^{-w}$ denotes the number of instances of $f$ currently associated with label $z_i^T$, and $c(\cdot)_{z_i^T}^{-w}$ the number of instances of any item currently associated with label $z_i^T$ (both terms exclude any instances contributed by $w$); $c(f)_\sigma$ signifies the number of instances of $f$ in element $\sigma$. For simplicity, I write $\lambda_m^l$ for $w_{1,\ldots,m}$ and $\lambda_m^r$ for $w_{m+1,\ldots,|w|}$. Once new values of $m, z_i^T, z_j^D$ are chosen for word $w$, counts for the features in $\lambda_m^l$ (if $m \neq \emptyset$) and $\lambda_m^r$ can be allocated to $z_j^D$ and $z_i^T$, respectively.

### Priors on morphological segmentation

For this paper's most basic inference procedures, the prior over morpheme boundaries is symmetric, with equal probability allocated to all possible segmentations of word $w$. In certain inference regimes, I employ one of two priors on segmentation incorporating the principle of Minimum Description Length, popular in unsupervised morphological segmentation (Goldsmith, 2001; Creutz & Lagus, 2007); these priors favor the insertion of morpheme boundaries which minimize the length of the code that generates the data. There are a number of ways to interpret this principle. Most intuitively, the "code" can be construed as the list of morph types, or alternatively, the sum of the lengths of morph types. Hence, an MDL or exponential prior on morphological segmentations disfavors analyses that add to the list, or the sum of (string) lengths of types in the list.

The first prior (MDL1), designed to keep the list of analyzed morphs short, assigns probability to a morphological segmentation for word $w$ proportional to the inverse of the number of morph types as currently analyzed, including the proposed segmentation for $w$; under the second approach (MDL2), the prior probability is inversely proportional to the sum of lengths of current morph types. I have employed these priors due to the importance of MDL in the literature on unsupervised segmentation, but remain somewhat skeptical as to whether HU numeral morphology can be rendered compact in the same manner as the morphology traditionally analyzed with MDL priors (e.g., of English, Finnish, Turkish, etc.), given the noisy allomorphy seen.

### Priors on cluster membership

Readers may note that the above formulae depart from traditional Dirichlet-Multinomial mixture models in that the Chinese Restaurant Process prior (a rich-get-richer scheme) over cluster membership is excluded. This prior, which makes it more likely for an item to be assigned to a cluster that already has many data points, seems inappropriate for this paper's model, which iterates over one token of each number, and should learn classes of roughly equal size. In one sampling regime, I place an exponential prior on TENS and DIGITS label membership, inversely proportional to the number of items currently assigned to the label in question (plus a concentration parameter). The intention here is to introduce a pressure toward clusters of uniform size.

## Inference procedure

Inference is carried out via Markov chain Monte Carlo. I run different versions of the model on three chains for 10000 iterations, discarding the first half of samples as burn-in. Each chain is initialized by randomly segmenting and assigning each item to a TENS and DIGITS label. Parameters are updated via Gibbs Sampling; for each number in 10–99, a morphological segmentation $m$, a TENS label $z^T$ and if relevant, a DIGITS label $z^D$ are drawn conditional on the labels currently assigned to all other data points (see eqq. 1–2). I use a simulated annealing procedure, raising each vector of update probabilities to the power of a constant $\frac{1}{\gamma}$, with $\gamma$ decreasing from 10 to 1 over the course of the burn-in. Code can be found at `github.com/chundrac/HUnumerals`.

I carry out an inference procedure using only bigrams as a phonological feature representation (2g); this is followed by a regime using unigrams and bigrams (1+2g). I modify the 1+2g procedure to incorporate an MDL prior sensitive to the length of the current list of morph types (MDL1), followed by an MDL prior sensitive to the sum of their lengths (MDL2). I attempted to see how the MDL1 prior (which showed better performance) affected the bigram model. Additionally, I ran a simulation which augmented the 1+2g/MDL1 model with a prior over component membership designed to keep clusters uniform (denoted by U).[1]

## Results

I use the overall F-measure (Fung et al., 2003) and the V-measure (Rosenberg & Hirschberg, 2007), two evaluation metrics designed to quantify the similarity between two classifications, in order to monitor convergence and measure overall accuracy (convergence was also assessed via chain log-likelihoods). I compute pairwise F- and V-measures between the maximum a posteriori (MAP) configuration of each chain to assess the degree to which chains return the same classification, interpreting values greater than .9 as a token of convergence between two chains. I evaluate each chain's accuracy by computing the F- and V-measures between the chain's MAP configuration and the true classification of the numbers. These values are found in Table 2. In general, MDL priors do not appear to improve inference for bigrams, and do not significantly improve inference for 1+2grams.

Table 3 displays the MAP configuration for the top chain (2) in the regime with highest overall accuracy (1+2g/MDL1/U). To measure the ACCURACY with which this regime decomposes individual numbers, I calculate the F-scores for each number's MAP TENS and DIGITS classifications with respect to its true TENS and DIGITS classifications, averaging these values. The resulting values are then averaged across chains. I calculate POSTERIOR UNCERTAINTY

[1] I also experimented with a procedure that excluded any TENS/DIGITS pairs from the proposal distribution for a given form that were assigned to any previous forms within a window of arbitrarily chosen size. However, this exacerbated the label-switching problem (a trivial issue); less trivially, it was difficult to motivate a window size which plausibly paralleled working memory.

by averaging the entropy of the posterior sample (comprising blocked draws of $m, z^T, z^D$) for each chain.

I extract numeral frequencies from the EMILLE Hindi Webnews corpus (Baker et al., 2002). For each number, accuracy and posterior uncertainty are plotted according to frequency in Figure 2, along with correlation coefficients and $p$-values. Both correlations are significant (albeit noisy), providing support for the idea that the HU numbers can be processed via a dual-route model. As seen in the lefthand plot, the majority of HU numbers occupy a quasi-Pareto frontier, indicating an efficient trade-off between decomposability and frequency. Several numbers in the teens (seen in the upper righthand corner of the plot) are both highly frequent and decomposable. These outliers in no way contradict the dual-route model, since a form's decomposability does not preclude the possibility that it is stored whole. However, a handful of numbers are found beneath the frontier (near the lower lefthand corner), meaning that they are both relatively infrequent and difficult to parse. These items can be viewed as vulnerable points in the grammar of HU numbers, and may be prone to "leakage" or analogical restructuring.

Table 2: F-/V-measures for different inference regimes

| | TEN | DIG | TEN | DIG | TEN | DIG |
|---|---|---|---|---|---|---|
| convergence | chain 1–2 | | chain 1–3 | | chain 2–3 | |
| 2g | .88/.87 | .77/.74 | .86/.86 | .81/.78 | .92/.91 | .95/.93 |
| 2g,/MDL1 | .77/.80 | .86/.82 | .78/.81 | .85/.84 | .87/.84 | .90/.88 |
| 1+2g | .88/.86 | .89/.88 | .90/.88 | .90/.89 | .99/.98 | .99/.99 |
| 1+2g/MDL1 | .93/.89 | .95/.95 | .94/.91 | .95/.95 | .99/.98 | 1/1 |
| 1+2g/MDL2 | .94/.92 | .95/.94 | .94/.92 | .95/.94 | 1/1 | 1/1 |
| 1+2g/MDL1/U | .88/.87 | .92/.92 | .89/.89 | .92/.92 | .97/.96 | 1/1 |
| over. accuracy | chain 1 | | chain 2 | | chain 3 | |
| 2g | .81/.81 | .76/.74 | .82/.84 | .86/.84 | .87/.88 | .92/.89 |
| 2g/MDL1 | .78/.79 | .82/.80 | .80/.82 | .89/.88 | .80/.81 | .85/.83 |
| 1+2g | .87/.86 | .89/.87 | .91/.91 | .90/.88 | .91/.91 | .92/.89 |
| 1+2g/MDL1 | .90/.89 | .88/.86 | .91/.91 | .9/.88 | .91/.91 | .9/.88 |
| 1+2g/MDL2 | .88/.87 | .90/.88 | .91/.91 | .9/.87 | .91/.91 | .9/.87 |
| 1+2g/MDL1/U | .91/.89 | .9/.9 | .93/.91 | .92/.89 | .91/.9 | .92/.89 |

Table 3: MAP configuration for 1+2g/MDL1/U, chain 2. Rows represent tens classification; columns represent digits classification. Numbers are represented by cardinality for readability. Asterisks $(*)$ mark numbers where the numerical representation TENS$\{i\}$, DIGITS$\{9\}$ maps to the representation TENS$\{i+1\}$, DIGITS$\{-1\}$

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 35 | | 34 | 31 | | 29∗ | 32 | 36 | 38 | | 30 |
| 75 | 77, 70 | 74 | 71 | 73 | 69∗ | 72 | 76 | 78 | | |
| 15 | 17, 16 | 14 | | 13 | | 12 | | 18 | 11 | 10 |
| 65 | 67 | 64 | 61 | 63 | 59∗ | 62 | 66 | 68 | | 60 |
| 44, 45 | 47, 27 | 40 | 41 | 43 | 39∗, 49∗ | 42 | 46 | 28, 48 | | |
| 25 | 37 | 24 | 21 | 33, 23 | 19∗ | 22 | 26 | | | 20 |
| 95 | 97 | 94 | 91 | 93 | | 92 | 96 | 98 | 90, 99 | |
| 55 | 57 | 54 | 51 | 53 | | 52 | 56 | 58 | | |
| 50, 85 | 87 | 84 | 81 | 83 | 79∗ | 82 | 86 | 88, 80 | 89 | |

## Discussion

The models presented in this paper show that although HU numerals 10–99 are morphologically irregular, a large number can be classified according to their component parts. However, quite a few forms are difficult to decompose, most of them of low magnitude and high frequency. In general, the models handled some types of allomorphy well, and others

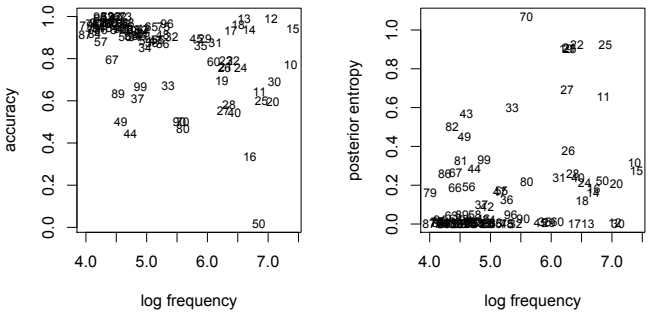


Figure 2: Log frequency as a predictor of model accuracy ($\rho = -.55, p < .001$) and post. uncertainty ($\rho = .38, p < .001$)

poorly. Forms containing the TENS$\{5\}$ allomorphs /-pən/ $\sim$ /-ʋən/ are grouped together, due to their agreement in two out of three segments. Surprisingly, /solə/ '16' was recognized as a member of the teens, despite the unique allomorph /-lə/; however, the model failed to properly classify it according to its digits place. Other forms with highly suppletive allomorphy (e.g., /ʊncɑs/ '49') were misclassified. Additionally, many simplex forms were not analyzed as monomorphemic, unless only a monomorphemic analysis was permitted under the proposal distribution.

As stated above, my results show that the HU numeral system's design is largely compatible with a dual-route model of access. In general, high-frequency items were more difficult for a computational model to decompose, indicating greater opacity. (Berger shows that many of these numbers were historically subject to erosion and evidently resistant to analogical changes that would otherwise make them more transparent and perceptually distinct.) At the same time, there are exceptions to this generalization: certain high-frequency items in the teens showed high accuracy, though this does not rule out the possibility that they are stored whole. Additionally, some problematic items are more opaque than would be expected, given their low frequency. It is likely that such vulnerable forms cause problems in planning and production.

The EMILLE Spoken Hindi corpus contains intriguing numeral variants (e.g., /iʈʰjɑnvɛ/ '91' by speaker ehinsp041, /sɪntɪjɑnvɛ/ '97' by ehinsp035, /ʊnɑnvɛ/ '89' by ehinsp044), though the data are too sparse to serve as the basis of a rigorous quantitative study. Many numbers are missing in the corpus; furthermore, the variation observed may stem from sources other than production difficulty, including transcriber error, multilingualism (with another Indic language; for example, speaker ehinsp017 utters the form /bɑvis/ '22', standard in Marathi but not HU), and stylistic factors. Studies of variability in the production of HU numerals — either in experimental contexts or naturalistic speech — will serve as a valuable research direction, particularly with an eye to whether vulnerable forms (i.e., sub-

optimal forms with higher opacity than expected relative to frequency) are subject to greater instability.

## Conclusion

In this paper, I have employed a simple and somewhat crude model of allomorphy, inspired in part by bag-of-words models used in document classification and intended to serve as a baseline for future work. A goal of this study was to test the limits of a simple mixture model in a HU numerical recognition task. A more sophisticated model of phonological processes may relate potential allomorphs to each other in terms of edits, as has been done in some MDL approaches (Virpioja et al., 2010). However, while such models can contend with or recover relatively regular allomorphy, no model has been designed, to my knowledge, to capture the highly noisy allomorphy found in the HU numeral system.

A true test of any computational model's value is in how well it agrees with human performance. A future direction for this work will involve carrying out experimental research to see how HU speakers process and produce numerical forms. It will serve us well to see how model inaccuracy fares as a predictor of greater response latency in psycholinguistic tasks. A joint approach which considers limitations in both experimental performance and computational simulation will help us identify weak points in this and other complex morphological systems that can potentially (though not obligatorily) undergo analogical change.

I have shown that frequency may facilitate the processing of more opaque HU numbers, but the question remains as to why most Indic number systems are on average more irregular than exact number systems found in other languages. Sociocultural factors may be partially responsible,[2] and their role in shaping cross-lingustic number systems should be taken into account along with that of functional need (cf. Xu & Regier, 2014).

## Acknowledgements

I am grateful to Stephan Meylan, Terry Regier, and three anonymous reviewers for helpful comments. All errors and infelicities are my own responsibility.

## References

Baayen, R. H. (1993). On frequency, transparency and productivity. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology 1992* (p. 181-208). Dordrecht: Kluwer.

Baker, P., Hardie, A., McEnery, T., Cunningham, H., & Gaizauskas, R. (2002). EMILLE, a 67-million word corpus of Indic languages: Data collection, mark-up and harmonisation. In *Proc. LREC 2002* (p. 819-825).

Berger, H. (1992). Modern Indo-Aryan. In J. Gvozdanović (Ed.), *Indo-European numerals* (Vol. 57, p. 243-287). Berlin: Walter de Gruyter.

Bright, W. (1969). Hindi numerals. *Working Papers in Linguistics (University of Hawaii)*, *9*, 29-47.

Brysbaert, M. (2005). Number recognition in different formats. In J. I. Campbell (Ed.), *Handbook of mathematical cognition* (p. 23-42). New York, Hove: Psychology Press.

Comrie, B. (n.d.). *Typology of numeral systems.* Available at `https://mpi-lingweb.shh.mpg.de/numeral/TypNumCuhk_11ho.pdf`. Accessed 1 October 2016.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, *4*, 1-34.

Fung, B. C. M., Wang, K., & Ester, M. (2003). Hierarchical document clustering using frequent itemsets. In *Proceedings of the SIAM International Conference on Data Mining.*

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, *27*(2), 153-198.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language and Linguistic Theory*, *30*, 859-896.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, *31*(2), 190-222.

Hurford, J. R. (1987). *Language and number: The emergence of a cognitive system*. Oxford: Blackwell.

Rosenberg, A., & Hirschberg, J. (2007). V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (p. 410-20). Prague: Association for Computational Linguistics.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606-21.

Turner, R. L. (1962–1966). *A comparative dictionary of the Indo-Aryan languages*. London: Oxford University Press.

Virpioja, S., Kohonen, O., & Lagus, K. (2010). Unsupervised morpheme analysis with Allomorfessor. In C. Peters et al. (Ed.), *CLEF 2009 Workshop, Part I* (Vol. 6241, p. 609-616). Heidelberg: Springer.

Xu, Y., & Regier, T. (2014). Numeral systems across languages support efficient communication: From approximate numerosity to recursion. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual meeting of the Cognitive Science Society.*

Yin, J., & Wang, J. (2014). A Dirichlet Multinomial Mixture Model-based approach for short text clustering. In *KDD '14: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (p. 233-242). New York: ACM.

[2] It is well known that South Asia is home to rigid societal hierarchies, and historically, exact number systems may have been the preserve of elite groups, while marginal groups relied on an alternative system (prior to language standardization). While this theory is blatant speculation, it is worth briefly noting that Sinhala, an Indic language whose speakers chiefly practice Buddhism (which preaches a doctrine of egalitarianism), developed a transparent number system.