Assigning Greammatical Gender to loanwords in Hindi-Urdu

Bhavya Pant
bhavyapant@umass.edu

Abstract

The question I'm interested in exploring is how speakers of languages with grammatical gender learn what gender a given word belongs to, especially when that word is borrowed from another language. Specifically, in this paper, I investigate how native speakers of Hindi-Urdu assign grammatical gender to words borrowed from English. What makes the Hindi-English case particularly interesting is that unlike Hindi, English does not assign grammatical gender to its nouns. Yet, Hindi speakers easily assign words borrowed from English to one gender category or another. This cross-linguistic gender-assignment has been attributed to a few different speaker strategies in the existing loanword typology literature. One such strategy emphasizes the role of morphology as a predictor of gender. To put this strategy to test, I model the gender-assignment process of English loanwords in Hindi as a function of the phonological similarity between a given loanword and the exiting words in the Hindi lexicon.

1 Introduction

Traditionally, when borrowing words from a language that encodes grammatical gender of its own, speakers of the borrowing languages simply inherit the gender of the borrowed words from their language of origin. Stolz (2009) demonstrates this for Italian words borrowed into Maltese and terms this strategy of inheriting gender directly from the source language as 'Gender Copy.' Nouns in Hindi all fall into one of two gender categories – masculine or feminine and are inflected for both number and case. English, however, does not assign grammatical gender to its nouns, and this eliminates Gender Copy as a viable strategy for Hindi speakers borrowing words from English. Yet, Hindi speakers, like those of other gendered languages have few problems assigning new or even artificial English nouns to one gender category or another (Corbett 1991). Moreover, the heavily code-mixed nature of modern colloquial Hindi (often termed as 'Hinglish') seems to suggest that this gender-assignment process occurs frequently and rapidly.

So, if not through cues from the source language, how are Hindi speakers assigning gender to words borrowed from English? The topic of gender assignment is somewhat contested within the loanword typology literature, and a few different speaker strategies have been proposed.

1.1 Strategy I

The first is that a borrowed word is simply assigned the gender of the equivalent word in the native language that it replaces (Singh et al., 2010). I find this theory to be not entirely convincing for a couple of reasons. Firstly, there are notable exceptions to this purported rule. For instance, the word for fork in Hindi (*kaanta*) is masculine, whereas its English equivalent 'fork' when borrowed into Hindi is treated as feminine. If loanwords were simply being assigned the gender of their equivalent form in the borrowing language, then such an asymmetry would not have existed. Moreover, not all borrowed English words necessarily have equivalent Hindi forms to map back to. This is especially true for technical terms (like *email*, *computer*) and entire concepts borrowed from the English lexicon (like *soda.*) Furthermore, even if the equivalent form of a borrowed word exists in Hindi, it might not be part of a native Hindi speaker's active vocabulary. This is because bilingual lexicons are not always equibiased i.e. the active vocabulary of a bilingual Hindi-English speaker will always not contain lemmas for all possible words in both languages. For instance, even though there exists a Hindi coun-

terpart to the word 'telephone,' it has fallen so drastically out of use that few Hindi speakers maintain it in their active vocabulary. So, the idea that speakers scan their lexicon for the Hindi equivalent of every English word they wish to borrow seems unlikely, since there isn't always a one-to-one mapping upon which to rely.

1.2 Strategy II

Another commonly proposed speaker strategy emphasizes the role of morphology in predicting the gender of borrowed words. According to this view, speakers pay special attention to certain (word-final) segments of borrowed words to predict their gender. These borrowed words are then assigned the gender most commonly associated with words in the native language that have a similar morphophonological makeup.

The vowel-endings of nouns in Hindi can often be good predictors of their gender, with nouns ending in /ɑː/ being generally masculine and those ending in /iː/ being generally feminine. This pattern is most clearly reflected in kinship terms such as grandfather (*dada*) and grandmother (*dadi*). When it comes to inanimate objects, however, the picture is less clear as there are many more exceptions. For instance, the word for water (*paani*) is masculine whereas that for garland (*maala*) and hope (*aasha*) is feminine. Even so, vowel-endings are likely a good place to start looking for gender-assignment cues of words borrowed from English.

To put the abovementioned strategy (Strategy II) to test, I modeled the gender-assignment process of English words in Hindi as a function of phonological similarity between the loanword and existing words in the Hindi lexicon.

2 Methods

To do this, I implemented a version of the generalized context analogical model proposed by Albright & Hayes (2003). This model defines similarity between two words in terms of their minimum edit (Levenstein) distance *d*.

$$similarity = \eta = e^{-\left(\frac{d}{s}\right)^p}$$

The minimum edit distance between two strings is the minimum number of editing operations needed to transform one string to the other. Each operation (insertion, deletion or substitution) incurs the transformation a certain penalty. Insertion and deletion operations incur a penalty of one point each, whereas substitution yields a penalty of two points, unless the substrings (in this case phonemes) being substituted are identical. For the purpose of this model, I modified the minimum edit distance so as to penalize more heavily any changes made the last *n* phonemes of the two words. Different values assigned to the parameter *n* made the cost of a particular transformation more or less sensitive to misalignment in word-endings. I also modified the substitution operation such that swapping two phonemes belonging to the consonant class was free of cost i.e. it did not incur the transformation any penalty. This meant that mismatching vowel-endings between the two words made the transformation between them most expensive.

The model was trained on data acquired from the Hindi-Urdu Treebank (HUTB) project, which is a multi-layered treebank corpus developed for Hindi and Urdu. From this corpus, I extracted 6000 Devanagari tokens along with their corresponding phonological transcriptions. The specific encoding scheme used to store these phonological forms was the WX-Roman Alphabetic Coding Scheme for Indian Languages, which is a transliteration scheme for representing Indian languages in ASCII. Every consonant and vowel in this scheme is mapped onto a single Roman alphabet, which is advantageous from a computational standpoint.

The resulting data was broken down into training, development and test sets. The development dataset was then used to ascertain the optimal value for the variable *s* in the similarity equation, which is an indication of how sensitive the model's similarity judgments are. When *s* is small, the model tends to rely primarily on a small set of very similar forms in forming its judgments. As *s* increases, the model becomes more sensitive to a broader range of forms. (Albright & Hayes, 2003) The optimal value of *s* that resulted in the highest accuracy for our model was *s=0.8.* For the sake of simplicity, the variable *p* in the similarity equation was set to 1, which is the value Albright & Hayes found to work best for their model. The test data which comprised of English words in their WX-form was similarly acquired from the treebank corpus.

For each loanword, I computed its similarity to Hindi nouns in either gender category (male or female). The similarity of a given word *w* to a col-

lection of words was simply defined as the sum of the pairwise similarities of *w* to all words in that collection. The gender predicted by my model was the one whose nouns have the highest overall similarity with the given loanword.

The accuracy of the model was calculated as the proportion of words in the test data whose actual (assigned) gender matched the gender predicted by our model. This calculation was performed over several values of *n* to bias the model towards prioritizing different sequences of sounds and the results were then tabulated.

3 Results

Model Variant	**Accuracy**
Albright & Hayes	74
Nearest neighbor	68
Prioritize last sound	66
Prioritize last 2 sounds	67
Prioritize last 3 sounds	65

Table 1: Accuracy of model variants

The model makes imperfect predictions with a peak accuracy of about 75 percent in the unbiased variant, which does not prioritize edits made to certain subsection of words over others. This is a surprising find since the prevailing belief is that speakers pay special attention to word-final segments to inform their gender-assignment decisions. Once certain segments are assigned more weight than others, this accuracy only goes down, peaking at about 67 percent for *n=2* (prioritizing last 2 sounds). The accuracy while prioritizing the very last sound (*n=1*) comes out to be 66 percent, while that for the last 3 sounds (*n=3*) comes out to be 65 percent.

To ensure that this inadequate performance is not just a result of the variable lengths of the words, with transformations between words of greatly differing lengths being penalized more severely than those between words of more similar lengths, I created a variant of the model which assigns a loanword its gender based solely on the gender of its nearest neighbor. Here, nearest neighbor is defined as the word in the Hindi lexicon which has the shortest (smallest) minimum edit distance to the loanword in question. This approach does not rely on the overall similarity of a loanword to an entire class (gender) of nouns, rather finds the word most similar to it (no matter what class it belongs to) and assigns it that gender. However, this approach too did not yield very satisfactory results, and resulted in an overall accuracy of just 68 percent.

These results are not too far off from those observed by Stolz (2007) in which she modelled the gender-assignment of English loanwords borrowed into Maltese as a function of their morpho-phonological form. There too, she found that analogy can account for only about 78 percent of the gender-assignment data.

Interestingly, upon examining the words most commonly misgendered by my model, I noticed an interesting pattern. For many of these words, their actual gender, as opposed to the one predicted by analogy, was the gender associated with their semantic equivalent in Hindi. For example, the borrowed word ‘shirt’ like most consonant-final words in Hindi is predicted to be masculine, when in reality it is treated by native Hindi speakers as masculine. What is notable here is that the equivalent word for ‘shirt’ in Hindi, namely *kamiz* is also feminine. Similarly, for ‘table’ (*mez*), ‘book’ (*kitab*), chair(*kursi*), ring(*angoothi*) and building (*imaarat*). In all these examples (and more), the actual gender of the words aligns less with its phonological neighbors and more with the gender of their equivalent forms in the borrowing language.

4 Discussion and Future Work

The results clearly demonstrate that analogy alone cannot account for the gender-assignment of loanwords in Hindi. Moreover, a closer look at the words most commonly misgendered by my model seems to suggest that Strategy I cannot yet be entirely ruled out as a potential speaker strategy. Maybe gender-mapping (which appears to be a more rule-based approach) is working in conjunction with analogy to predict the gender of borrowed words. My intuition regarding this is that given a loanword, speakers first attempt to find an equivalent form of that word in their native language from which to borrow its gender, and if that fails, then they resort to analogy (form) as a second measure. Of course, I don’t expect this process to be perfect, but it might be able to explain some of the asymmetries we see in my results.

One way to tease apart the two strategies mentioned above would be to perform the same experiment with nonce forms i.e. train the model on

Hindi data and then have it make predictions regarding the gender of made-up words. It would be interesting compare the performance of this model to that of humans. This condition is especially worth exploring since unlike for actual English words, nonce words don't have an equivalent form in the native language to map back to, and so humans must rely entirely on analogy to assign these made-up words their gender.

Another strategy worth exploring, is what Stolz (2007) terms 'Borrowing Competition'. According to this strategy, in the absence of an equivalent of the loanword in the native language to map back to, speakers tend to assign to it the gender of its nearest lexical equivalent. One could put this strategy to test, both in isolation as well as in conjunction with the two strategies mentioned above by making use of a semantic word nets, which is a lexical database of words grouped into sets of synonyms called synsets, which are all interconnected by means of semantic relations.

One such wordnet for the Hindi language, developed by The Indian The Indian Institute of Technology, Bombay is a potential useful resource to get started with this task.

5 Acknowledgements

I would like to thank Professor Gaja Jarosz and Max Nelson for all their help and feedback, Professor Rajesh Bhatt for his help with finding the data, and my peers in the Fall 2019 Cognitive Modelling course for their helpful comments and suggestions.

References

Stolz, Christel. 2009. *A different kind of gender problem: Maltese loan-word gender from a typological perspective.*

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press

Singh, Smriti, Vaijayanthi M. Sarma, and Stefan Muller. 2010. *Hindi Noun Inflection and Distributed Morphology.*

Albright A., & Hayes B. (2003). *Rules vs. analogy in English past tenses: A computational/experimental study*. Cognition, 90, 119–161.