English Influence on Gujarati: A Statistical Examination of News Headlines

Prina Doshi1

1Carnegie Mellon University

phdoshi@andrew.cmu.edu

Abstract

When cultures come into contact with one another, their languages may not always be similar.  Many times, foreign words must be incorporated into a new language, especially in instances when no direct translation can be found. This study examines what English words are borrowed into Gujarati. Using a public dataset of headlines taken from 3 categories (technology, entertainment, and business) of the News18 Gujarati Website, a database of headlines with 5026 total characters was built. A human rater classified each word as borrowed. These words were then manually classified by their part of speech (noun, adjective/adverb, verb). We found that there was no difference between the number of borrowed words across the three headline categories or the three parts of speech. There was also no relationship between part of speech and headline category in terms of number of borrowed English words.

**Index Terms**: transliteration, borrowing, loanwords, Gujarati

# Introduction

## Background on Gujarati

Gujarati is an Indo-European language spoken by approximately 55 million speakers around the world [1]. Although it is primarily spoken in the state of Gujarat, India, the diaspora of Gujarati speakers has spread to other countries, including the United States, United Kingdom, and Singapore [2]. In the United States, Gujarati was one of the fastest growing foreign language-groups between 2010 and 2017 [3].

Gujarati uses a script modified from the Devanāgarī script, which is used in languages like Hindi, Marathi, and Sanskrit. The primary difference between the two is the adjustment of some characters and the removal of the horizontal line across the top of characters.

गुजराती (1)

ગુજરાતી (2)

The word ‘Gujarati’ is written in both Hindi script (1) and Gujarati script (2). The main differences can be seen in the character that creates the ‘ja’ sound. In Gujarati, this is created with ‘જ’ and in Hindi it is created using the ‘ज’ character. Gujarati is an Abugida script, indicating that consonants are built from characters, and vowel sounds are made using modifications to these characters.

Despite being somewhat known for incorrect or absurd pronunciations of English words, a 2011 census in India showed that 13.85% of those living in Gujarat can speak English, which was higher than the national average [4]. Despite this, little research has been done on the interaction between Gujarati and English.

## Lexical Borrowing

When cultures interact with each other, there is a natural exchange of concepts that may result in a transaction occurring between the two languages. When a word is exchanged from a foreign (donor) language into the native one, it can be considered an instance of borrowing. This study takes its definition of loanwords from [5, pp. 43-44], who state that “[l]inguists identify words as loanwords if they have a shape and meaning that is very similar to the shape and meaning of a word from another language from which it could have been taken. . . and if the similarities have no plausible alternative explanation.” For instance, the Hindi word ‘chutney’ and the Spanish word ‘taco’ were both borrowed into English.

If the resemblance between the words has a reasonable explanation, such as having a common language family or root, it would not be considered an instance of borrowing. For example, in Gujarati, the word ‘નામ’ is pronounced as ‘nāma’ and could be considered similar to the English ‘name’ in terms of both pronunciation and meaning. However, both words have a common root in this instance and therefore are not considered instances of borrowing.

Prior work regarding lexical borrowing has categorized it into two primary types. The first is called cultural borrowing, which occurs when a new concept from a different culture must be conveyed into the native one [5]. The other is called an instance of core borrowing, and it occurs when the native language has a duplicate concept with a new word. [6] first made the distinction between types of borrowed words, with the assumption that ‘cultural borrowings’ occur when a word is a foreign concept that does not exist in the original culture. For example, when ‘taco’ is borrowed from Spanish, this word did not exist in the native language since it is a new food.

Although cultural borrowings need not occur, they may be considered the more understandable of the two because foreign concepts may be introduced and need a word very quickly. In many instances, cultures have come up with their own word for a foreign concept rather than borrowing the word from another culture. If a word is already well known in another language, it becomes more convenient to use it rather than create a native-language alternative [5]. The predominant theory about how a cultural borrowing becomes popular is that bilingual speakers begin using it, eventually causing other speakers to do the same. In Gujarati, some examples of cultural borrowings may include examples of modern technology, such as ‘telephone.’

Conversely, there are conflicting theories as to why core borrowings occur. [6] states that a core borrowing occurs after code-switching, when a foreign word is interjected into the native language within the middle of a sentence. Others ([5], [7]) also claim that core borrowings primarily happen due to the ‘prestige’ associated with speaking in the donor language. Since bilingualism and knowledge of a certain foreign language may be associated with access to an education, speaking in the foreign language rather than the native one may be a method for demonstrating socioeconomic status. In particular, [6] argues that core borrowings occur much more gradually than cultural borrowings, which can be very sudden.

[8], cited by [9], found that nouns are more likely to be borrowed than other parts of speech, followed by adjectives and verbs. [9] finds, in a study about English instances of borrowing, that nouns were borrowed at the highest rate, followed by verbs, adjectives. They find similar results for Dutch loanwords from English, with nouns borrowed the most, followed by adjectives and verbs. This leads to the understanding that parts of speech impact the rate of borrowing for words. [9] also found that abstract words were more likely to be borrowed, and that frequency of use of the word is related to borrowing.

## Research Question

Many projects have attempted to understand the process of borrowing for languages like Chinese [10], Dutch [9], and French [11]. One project to document loanwords is the World Loanword Database, which categorizes loanwords from different languages [5]. However, there is little information on English loanwords to the language of Gujarati.

Since English is considered a lingua franca for business and technology, it may have been easier for other cultures such as Gujarati to simply adopt these words rather than create new alternatives that must be translated each time. For this instance, we hypothesize that there may be differences in the amount of borrowing seen in different types of Gujarati headlines, since some fields like technology may have more instances of cultural borrowing. This leads us to the first research question: Is there a difference in the amount of borrowing found between different news headline types?

Considering that abstract concepts may be included more frequently within certain subjects, there may be reason to believe that both news category and part of speech may impact the amount of transliteration found. This leads us to the second research question: Is there a relationship between the part of speech of the borrowed word and the category of news it came from?

# Experiment

## Method

### Materials

For this study, data were taken from a free, public dataset [12] that web-scraped headlines from a Gujarati website, News18 Gujarati. These headlines were taken from three categories listed on the site: technology, entertainment, and business. From this large dataset, a database of 97 headlines (5026 characters) was created.

### Transliteration Process

For each headline, the words that were borrowed from English were noted as new observations. To identify these, each headline was reviewed individually. If a word had a perfect or near-perfect phonetic match in English, and if the intended meaning appeared to match the English word, it is considered an instance of borrowing.

મોદી સરકાર 1 **ફેબ્રુઆરીએ** વચગાળાનું **બજેટ** રજૂ કરશે, આ જાહેરાતો થઈ શકે (3)

Modi sarakaara 1 **phebruarie** vachagaalaanum **bajeta** raju karashe, aa jaaheraata tha-ee shake (4)

Modi government 1st **February** interim **budget** present they will, this announcement could be (5)

Modi government will present the interim **budget** on **February** 1st, this could be announced (6)

For example, the Gujarati headline (3) was first phonetically translated (4), with the bolded words identified as possible instances of transliteration. Based on the gloss (5) and translation (6), they were identified as loanwords borrowed from English. When proper nouns were used in the dataset, if they were originally English or non-Indian names, these were considered instances of borrowing. Although code-mixing also occurred in this data, these cases were not classified as instances of borrowing since they were written in the English script.

Once observations of borrowing were noted, these words were classified into their respective parts of speech (noun, proper noun, adjective, adverb, verb) by a human rater based on the headline they were found within. Within the headlines assessed, 189 observations of borrowed words were found.

## Results

To compare average counts of borrowed words in different headline categories, the pure count must be normalized by the length of the headline. The number of total words in each headline was counted and a new normalized borrowing count was created by dividing the number of borrowed words by the total number of words. Statistical analysis was conducted using R version 4.1.1

To test the first research question, an analysis of variance test was conducted between the normalized count of borrowed words among the three news headline. From Figure 1, we can see that there does not seem to be a significant difference in the average number of transliterated words between the three headline types. Although Entertainment seems to have a slightly higher average, overall this number appears to be between 0.15 and 0.20 for all three categories. This is supported by the descriptive statistics shown in Table 1 for each of the headline categories.

Chart, box and whisker chart

Description automatically generatedFigure 1. *Average Borrowed Words per Headline, by Category*

Table 1. *Descriptive Statistics by Headline Type.*

|  |  |  |
| --- | --- | --- |
| **Headline Category** | **Mean Normalized Count** | **SD** |
| Technology | 0.190 | 0.146 |
| Business | 0.178 | 0.161 |
| Entertainment | 0.203 | 0.159 |

The data are classified non-normal, as seen in Figure 2, and transformations on the data did not make the data normal. Therefore, the Kruskal-Wallis non-parametric test was used, with power of 0.99.

Chart, histogram

Description automatically generated

Figure 2. *Count of* *Average Number of Borrowed Words per Headline, by category.*

We found that there was no difference between the number of borrowed words across the three headline categories or the three parts of speech, (Kruskal-Wallis 𝜒2 (2, N = 97) = 0.875, p = 0.64). A Dunn-Test post-hoc analysis is seen in Table 2, showing that each comparison shows no difference. These results can be seen in Figure 2.

Table 2. *Dunn-Test Results.*

|  |  |  |
| --- | --- | --- |
| **Column - Row**  **(p-value)** | **Business** | **Entertainment** |
| **Entertainment** | -0.887666 (0.5621) |  |
| **Technology** | -0.667038 (0.7571) | 0.164715 (1.000) |

To test the second research question, a chi-squared test of independence with power 0.99 is used to assess whether there is a relationship between headline category and part of speech for each word. The contingency table is shown in Table 3. We find that there was no relationship between part of speech and headline category in terms of number of borrowed English words, (𝜒2(2, N = 189) = 1.058, p = 0.59, V = 0.08).

Table 3. *Contingency Table of Borrowed Words and Headline Category.*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Business** | **Technology** | **Entertainment** |
| **Noun, Proper Noun** | 50 | 40 | 56 |
| **Adjective, Adverb, Verb, Pronoun** | 12 | 15 | 15 |

# Discussion

## Findings and Limitations

Statistical analysis found no significant difference between the mean normalized counts of borrowed words between different headline categories. Although entertainment headlines had a slightly higher count, there is no true difference between the means. This is somewhat surprising since there are typically more instances of cultural borrowing in the technology categories compared to other fields. This could show that borrowing may occur in different ways, such as core borrowings, in the other categories.

Tests for the second research question also found that there is no relationship between part of speech borrowed and headline type. This is also surprising since typically nouns are borrowed more frequently, and categories like entertainment may be using more proper nouns for things like names. However, this was not the case, since no relationship was found. One reason for this may be the sample size. Due to the number of headlines analyzed, the parts of speech were combined into two larger categories in order to meet the requirements to run a chi-squared test. Larger sample sizes with more observations for each part of speech may yield results that highlight differences between the specific categories (adjectives, verbs, adverbs) that were lost in this study.

## Additional Observations

Although the results were not statistically significant, there were still some interesting patterns shown in the data. In some cases, words that had no Gujarati equivalent were borrowed, such as ‘February.’ Even though India typically follows a lunar calendar, sometimes months must be given in the Julian format. Conversely, a word such as ‘budget’ has an equivalent concept in Gujarati, but it is possible that even native speakers are more familiar with the English word rather than the original Gujarati one. The question of why ‘budget’ and other words may be borrowed could have something to do with influences of British occupation on India, and also may be due to increases in global communication and connection.

One other interesting pattern involved instances where sound and meaning borrowing occurred with a Gujarati morphological modification. For instance, the word ‘center’ was borrowed and written out in Gujarati script to sound like ‘senṭar.’ It was written in a plural form, but instead of using the English plural marker ‘s’, which would have been ‘centers,’ it was written using the Gujarati plural marker ‘o’ as ‘senṭaro.’ This may indicate that some words in these headlines underwent a two-step borrowing process first involving the English sound and meaning and then undergoing native Gujarati morphological rules.

Finally, one interesting phenomenon is that certain articles used code-switching in headlines, even to refer to English names, whereas others used a transliterated or borrowed word written in the Gujarati script to write the name. Understanding why this may occur could be beneficial to learning more about the process by which words become borrowed.

## Future Research

Future research may replicate this study and attempt to analyze a greater number of headlines in order to find specific differences in the parts of speech. It may also be valuable to analyze a different source, either from a new news website, a form or print media, or the words within the article. Furthermore, developing a computational algorithm to automatically categorize words as instances of borrowing may help assess a much larger dataset.

Another interesting approach to understanding why certain words have been borrowed would be to ask proficient speakers of both Gujarati and English to identify the Gujarati equivalent of each borrowed word. The time it takes for these speakers to recall this word would be noted and could be compared to another set of control group words. This would provide insight about whether or not the words are likely cultural or core borrowings. Knowing this could be beneficial for differentiating between borrowing for words like ‘February’ and ‘budget.’

# Conclusions

The goal of this study was to understand more about instances of borrowing from English to Gujarati. We found that there is no difference in the amount of borrowing observed in different news headline categories, and we also found no relationship between the part of speech of borrowed words and the headline category from which they were found. However, even with these results, we observed some interesting instances of lexical borrowing and modification that provide a better understanding of the interaction between English and Gujarati.

# Acknowledgements

The author would like to thank Dr. Seth Weiner for guidance on this project.

# References

1. *Gujarati: The language spoken by more than 55 million people*. (2017, January). The Straits Times. https://www.straitstimes.com/singapore/gujarati-the-language-spoken-by-more-than-55-million-people.
2. Britannica, T. Editors of Encyclopaedia (2020, January 24). Gujarati language. Encyclopedia Britannica. https://www.britannica.com/topic/Gujarati-language
3. Bhattacharya, A. (2018, September). *America’s fastest growing foreign language is from south India*. Quartz India. https://qz.com/india/1399825/telugu-is-uss-fastest-growing-foreign-language/
4. Khan, S. (2018, November). *In Gujarat, where money usually talks, English has gained currency.* The Times of India. https://timesofindia.indiatimes.com/city/ahmedabad/in-gujarat-where-money-usually-talks-english-has-gained-currency/articleshow/66613057.cms
5. Haspelmath, M. (2009). II. Lexical borrowing: Concepts and issues. In *Loanwords in the World's Languages: A Comparative Handbook* (pp. 35-54). Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110218442.35
6. Myers-Scotton, C. (2002). Contact Linguistics: Bilingual Encounters and Grammatical Outcomes. : Oxford University Press. Retrieved 10 Dec. 2021, from https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780198299530.001.0001/acprof-9780198299530.
7. Winford, D. (2010). Contact and Borrowing. In *The Handbook of Language Contact*, R. Hickey (Ed.). https://doi.org/10.1002/9781444318159.ch8
8. Cassidy, F. G. (1999). English: A Germanic Language? In G. F. Carr, W. Harbert, & L., Zhang (Eds.). Interdigitations: Essays for Irmengard Rauch (pp. 75-79). New York: Peter Lang.
9. Monaghan, P., & Roberts, S. G. (2019). Cognitive influences in language evolution: Psycholinguistic predictors of loan word borrowing. *Cognition, 186*, 147-158. https://doi.org/10.1016/j.cognition.2019.02.007
10. Cook, A. (2018). A typology of lexical borrowing in Modern Standard Chinese. *Lingua Sinica, 4*. https://doi.org/10.1186/s40655-018-0038-7.
11. Paradis, C., & LaCharité, D. (2008). Apparent phonetic approximation: English loanwords in Old Quebec French. *Journal of Linguistics, 44*(01), 87-128. http://dx.doi.org/10.1017/S0022226707004963.
12. Gaurav. (2019, December). Gujarati News Dataset, Version 1. Retrieved from https://www.kaggle.com/disisbig/gujarati-news-dataset.