Internet readiness of Assamese

A comparison with three languages

Satyajit Nath
https://www.linkedin.com/in/satnath/

Abstract

The effectiveness of tools and mechanisms available on computers and the internet for a given human language is an indicator of its *internet readiness*, a term defined in some detail in this paper. However, for Assamese - an Indic language used by over 15 million people, and one of 22 languages recognized by India's constitution to "grow rapidly in richness and become effective means of communicating modern knowledge"[1] - its internet readiness is currently lower than that of other major languages. Lower internet readiness makes it harder for people to use the language on the internet. It is important to improve this situation to enable people to create, find, and read Assamese content more easily on the internet - an important priority for the language today. This paper first defines a framework of required elements to analyze internet readiness of any human language. These elements are then used to compare and contrast the level of internet readiness of Assamese with three other languages, identifying areas that need attention for Assamese. The paper concludes with recommendations on priorities to increase the level of internet readiness of the language.

Introduction

Internet readiness of a human language can be thought of as the maturity and effectiveness of available tools and mechanisms on computers and the internet to display, write, print and process text in that language. Increased internet readiness improves the ability of users of the language to consume and generate content in the language. This acceleration of content generation on the internet is an important priority for Assamese today.

To analyze the internet readiness of Assamese, this article compares it with the internet readiness of other languages. For this purpose, three other languages were selected - English, Dutch and Bengali. English and Dutch are a relevant pair for this purpose because both use essentially the same script, and Dutch has, at least, an order of magnitude less speakers than English. Bengali and Assamese have a similar relationship with each other. Both use essentially the same script (with two additional letters for Assamese), and Assamese has an order of magnitude less speakers than Bengali. Thus, comparing the internet readiness of these four languages together provides useful insights.

To analyze the details of internet readiness, this article identifies ten different elements required to support a human language on computers and the internet. Each element is analyzed for each of

[1] https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India

the four languages with the goal to compare and contrast the level of support. Finally, using that analysis, a numeric score is provided for each language to quantify internet readiness of that language.

Assamese scored the lowest among the four languages for internet readiness, showing that there is a lot left to do for the language compared to the other three languages.

The rest of this article describes the detailed analysis used to compute the internet readiness scores. The article concludes by identifying areas of further work which need to be prioritized for Assamese to improve internet readiness.

The elements

For the purpose of this analysis, Table 1 below identifies ten elements required to support a human language on computers and the internet.

	Element needed on computers and the internet	**What human task does it support for the language?**
1.	Standard digital representation of all letters in the script	Storing/communicating text
2.	Ability to show letters of the language on screen/print	Reading text
3.	Ability to input letters of the language	Writing text
4.	Ability to check spelling of words input	Writing text
5.	Ability to check grammar of sentences input	Writing text
6.	A defined alphabetical order for the digital representation of text	Ordering texts alphabetically
7.	Ability to label text in web pages with the language	Finding text on the internet
8.	Ability to limit web-searches to the language	Finding text on the internet
9.	Optical Character Recognition to convert printed matter into the standard digital representation of all letters in the underlying text	Converting existing printed matter for reading, editing, searching on the internet

10.	A Romanization (or transliteration) standard for the language	Using Roman script on computers to easily write and read the language on the internet without learning the script

Table 1: Elements to support a human language on computers and the internet

Scoring criteria

For each of the ten elements identified in Table 1, the following criteria are used to provide a quantitative **element score** for each of the four languages.

- 1 point: Standard in place or full support for the element
- 0.75 points: Extensive readiness for the element with some gaps
- 0.5 points: Limited or Basic support for the element
- 0 points: No or unclear support for the element

Detailed analysis

This section provides detailed analysis of the readiness of the four languages in each of the ten elements defined in the previous section. Each subsection below focuses on a single element out of the ten.

1. Standard digital representation of all letters in the script

How does the software industry meet this need for a given human language?

Unicode specifies a unique numeric code for each written symbol used in the script for the language. The collection of character codes for a script is known as a CODE CHART.

Status for the four languages

- Standard in place for all four languages

Analysis

- English and Dutch use the Unicode code chart named Latin[23].
- Assamese and Bengali use the Unicode code chart named Bengali[1].
- The *dari* (।) (U+964) character used in Assamese is specified in the Unicode code chart named Devanagari[24], as is the *double dari* (॥) (U+965) character. These punctuation marks are for common use in the scripts of India despite being placed in the code chart named Devanagari.

Element Score

- 1 mark for all four languages

2. Ability to render letters of the language on screen/print

How does the software industry meet this need for a given human language?

Vendors develop "fonts" that contain digital representation of the actual shapes of the characters in various code charts. These shapes are called "glyphs". These glyphs are used to display or print the shapes of sequences of character codes in the script. The OpenType specification[2] is the authoritative standard for such fonts.

Status for the four languages

- Supported on all platforms out of the box for all four languages

Analysis

- English and Dutch use fonts that include glyphs for character codes in the Unicode Latin code chart. Assamese and Bengali use fonts that include glyphs for character codes in the Unicode code chart named Bengali. Note that a single font can support multiple code charts. So, it is very common to use a single font for all of English, Dutch, Bengali, and Assamese text.
- One issue to which font designers need to pay attention is that Assamese has two special ligatures:

  ৰু ৰূ

  The special shapes required for these Assamese ligatures are specified in the Unicode code chart named Bengali[1].
- These ligatures are composed of ৰ (U+9F0) with ু (U+989) and ূ (U+98A) respectively.
- A few fonts that support the Unicode code chart named Bengali include the correct glyphs for these Assamese ligatures (e.g. the third-party font Lohit Assamese[26] and the font Shonar Bangla included in Windows 10 , Garhgaya-Assamese, Noto Sarif/San Bengali etc.).
- But, many other fonts that support the code chart do not provide the correct ligatures for those Assamese characters. E.g. the font Vrinda in Windows 10 renders them incorrectly as

  

- Further, some fonts do not correctly render *extended* conjuncts that end with ৰু or ৰূ.
- E.g. the correct glyph for the conjunct composed of ভ (U+9AD), ৰ (U+9F0) and ু (U+9C1) is

  ভ্রু (U+9AD U+9CD U+9F0 U+9C1).

  However, some fonts on Windows render this conjunct wrongly as:

  

- These problems seen with some fonts in Windows 10 have not been seen in fonts included in the Android, macOS Catalina, and iOS13 platforms.

- The fonts that deviate from the correct rendering of the ligatures for these Assamese character sequences should be identified and the vendors should be contacted to update the glyphs to adhere to the requirements noted in the Unicode specification.

Element Score

- 1 mark for English, Dutch, and Bengali
- 0.75 marks for Assamese (due to the deficiencies of fonts on Windows as described in Analysis above)

3. Ability to input letters of the language

How does the software industry meet this need for a given human language?

Vendors develop "input method" software that takes user input for letters from keyboard and other input devices and converts to the appropriate Unicode character codes.

Status for the four languages

- Supported on all platforms out of the box for all four languages

Analysis

- Apple was lagging behind Microsoft and Google on this for a few years. But in 2019, iOS13 and macOS Catalina added native support for Assamese keyboard input. Now all the major platforms (Windows, Android, macOS, iPhone/iPad) have full support for Assamese keyboard input out of the box.
- Earlier, several third-party keyboards were indispensable tools till standard support arrived for Assamese on all platforms. Some notable examples are: xobdo.org keyboard[14, 15], Lachit[16], Luitpad[4], Google Indic[17], gBoard[18], Pramukh Typepad[19], Rodali[20].

Element Score

- 1 mark for all four languages

4. Ability to check spelling of words input

How does the software industry meet this need for a given human language?

Vendors develop "spell checker" software for the language that provides feedback to the user about spelling mistakes in the text they are inputting.

Status for the four languages

- Widely available for English, Dutch, and Bengali
- Only basic support for Assamese

Analysis

- Xobdo Litikai[3] introduced the much needed spell-check capability for Assamese in 2012. LuitPad[4] was also added in the same year.
- But integration of Assamese spell checkers into the standard text input fields of current operating systems is still missing, unlike for the other languages.
- This is important to ensure people can type efficiently and with correct spelling in Assamese, without having to stop to get the spelling checked in a tool separate from where they are typing the text.

Element Score

- 1 mark for English, Dutch, and Bengali
- 0.5 marks for Assamese

5. Ability to check grammar of sentences input

How does the software industry meet this need for a given human language?

Vendors develop "grammar checker" software for the language that provides feedback to the user about grammar mistakes in the text they are inputting.

Status for the four languages

- Widely available for English and Dutch
- Limited support for Bengali
- Not available yet for Assamese

Analysis

- This is an area where active work is needed for Assamese.

Element Score

- 1 mark for English and Dutch
- 0.5 marks for Bengali
- 0 marks for Assamese

6. A defined alphabetical order for the digital representation of text

How does the software industry meet this need for a given human language?

Unicode specifies a collation order as part of the LOCALE definition for the human language.

Status for the four languages

- Standard in place for all four languages

Analysis

- Alphabetical order used in the Assamese language is specified separately from that used in the Bengali language.
- The position of ক্ষ (U+995 U+9CD U+9B7) in the alphabetical order of Assamese (which is different than in Bengali) is correctly specified in the standard[5].
- The alphabetical order for Dutch is, similarly specified separately from English, with additional digraphs between **i** (U+69) and **j** (U+6A), that are unique to the order used in the Dutch language[21].
- Swedish, another language that shares the Latin script with Dutch and English, has an even more pronounced distinction in its alphabetical order, with three letters specified after **z** (U+7A), that are unique to the order used in the Swedish language[22].

Element Score

- 1 mark for all four languages

7. Ability to label text in web pages with the language

How does the software industry meet this need for a given human language?

ISO has defined standard language codes for languages around the world. W3C has defined tags to specify a language code in any web page to indicate the language of the text on that page.

Status for the four languages

- Standard in place for all four languages
- The language tag specified by W3C to use for Assamese web pages is '**as**'

Analysis

- W3C has provided guidance on how to use language tags in web pages[6].
- However, web pages written in the Assamese language are not being tagged according to the standard. This needs attention because it is a prerequisite to limit web searches to Assamese content.
- Note that Bengali language content is also not being tagged according to this standard.

Element Score

- 1 mark for English and Dutch
- 0.5 mark for Bengali and Assamese (supported but unused)

8. Ability to limit web-searches to the language

How does the software industry meet this need for a given human language?

Web search engines like Google have provided an 'Advanced Search' capability where a user can limit search results to specific language[7].

Status for the four languages

- There is support on Google for English and Dutch
- There is no support on Google for Bengali and Assamese

Analysis

- The absence of Bengali as a language choice in 'Advanced Search' does not impact searching for web pages written in that language.
- But the absence of Assamese has a huge impact. It makes it very difficult to search for web pages in that language.
- A separate article[8] by this author analyzes why that is true, describes a workaround - searching with ৰ (U+9F0) - and discusses some solutions.
- To fix the issue, it is important for Google to add support for both languages in 'Advanced Search' as well as in the list for its 'Search Language' setting[9].
- It is also important for content creators to tag their web pages as Assamese (as noted in Analysis of element 7 above).

Element Score

- 1 for English and Dutch
- 0 for Bengali and Assamese (but the Analysis above is important)

9. Optical Character Recognition to convert printed matter into the standard digital representation of all letters in the underlying text

How does the software industry meet this need for a given human language?

Vendors develop language-specific OCR software to perform this conversion from an image of printed matter to the standard digital representation of the underlying text in that image.

Status for the four languages

- Available for English, Dutch and Bengali
- Basic support for Assamese

Analysis

- Pramukh OCR[10] is a great tool available on Android to the public today for Assamese.
- But in addition to basic support to recognize the shape of the letters, it is important for OCR software to integrate with Assamese spelling and grammar checkers. This is important for increasing accuracy in the converted digital representation of the text.
- This is an area that needs a lot of attention for Assamese. There are some research projects in this space that could result in more tools.

Element Score

- 1 for English, Dutch and Bengali
- 0.5 for Assamese

10. A Romanization (or transliteration) standard for the language

How does the software industry meet this need for a given human language?

ISO has defined Romanization standards for a wide variety of languages[11].

Status for the four languages

- Not applicable to English and Dutch (script is Roman)
- Standard is in place for Bengali
- Not clear if there is a standard in place for Assamese

Analysis

- The lack of clarity about a standard for Assamese is due to the existence of a wiktionary web page documenting an Assamese transliteration table[12].
- This transliteration table is phonetically accurate for Assamese. Further, wiktionary notes on that web page that the table is based on ISO 15919[13] -- the standard produced by the ISO for transliteration for all Indic scripts.
- But, the current perception is that there is no transliteration standard for Assamese (and the one for Bengali is being offered as the standard, which does not work well for Assamese pronunciation of many letters).
- Clarification is needed whether wiktionary's claim is accurate that its transliteration table for Assamese is based on the ISO standard. To get confirmation, discussion is needed with wiktionary and the ISO. This discussion is required because ISO does not make its standard freely available in full with the necessary details to verify this independently.

Element Score

- Not applicable to English and Dutch (because the languages use the Roman script)
- 1 for Bengali
- 0 for Assamese (until we get clarification from wiktionary and ISO as noted in Analysis above)

Element Score Summary

The Element Scores for all four languages from the Detailed Analysis section above are summarized in Table 2 below.

	Element needed on computers and the internet	**English**	**Dutch**	**Bengali**	**Assamese**
1.	Standard digital representation of all letters in the script	1	1	1	1
2.	Ability to render letters of the language on screen/print	1	1	1	0.75

3.	Ability to input letters of the language	1	1	1	1
4.	Ability to check spelling of words input	1	1	1	0.5
5.	Ability to check grammar of sentences input	1	1	0.5	0
6.	A defined alphabetical order for the digital representation of text	1	1	1	1
7.	Ability to label text in web pages with the language	1	1	0.5	0.5
8.	Ability to limit web-searches to the language	1	1	0	0
9.	Optical Character Recognition to convert printed matter into the standard digital representation of all letters in the underlying text	1	1	1	0.5
10.	A Romanization (or transliteration) standard for the language	N/A	N/A	1	0
	Internet Readiness Score of the language	**9/9 100%**	**9/9 100%**	**8/10 80%**	**5.25/10 52.5%**

Table 2: Summary of scores for the ten elements for all four languages

Conclusion

The analysis shows that, compared to the other languages considered here, many elements of internet readiness of Assamese are missing or with low capability today.

Despite the low internet readiness so far, a respectable amount of Assamese content has already been created on the internet. An approximate count on Google[2] shows that there are currently about 2.2 million Assamese language web pages.This is a testament to the will and perseverance of many people who have worked tirelessly to generate that content on Assamese wikipedia, various other websites, personal blogs and through other means. A workaround is also worth mentioning to more-or-less limit web searches to Assamese web pages - simply by adding ৰ (U+9F0) to the search term (e.g. on Google, by using কবিতা ৰ to search for Assamese web pages about poetry). But we can only imagine how much more can be done, and at an accelerated pace, if internet readiness of the language is improved.

Figure 1 shows the elements and their readiness for Assamese in the technology stack used in computers and the internet. Unicode includes all the character codes needed for Assamese in the

[2] Searching for the letter ৰ (U+9F0)

code chart, and takes proper consideration for font requirements for the unique characters of Assamese, as analyzed here. Further, distinct language-specific attributes, like alphabetical order, are specified in a separate language-specific Unicode standard for Assamese, as also shown here. Most of the elements that are missing or with low capability are elsewhere in the stack as shown in Figure 1.

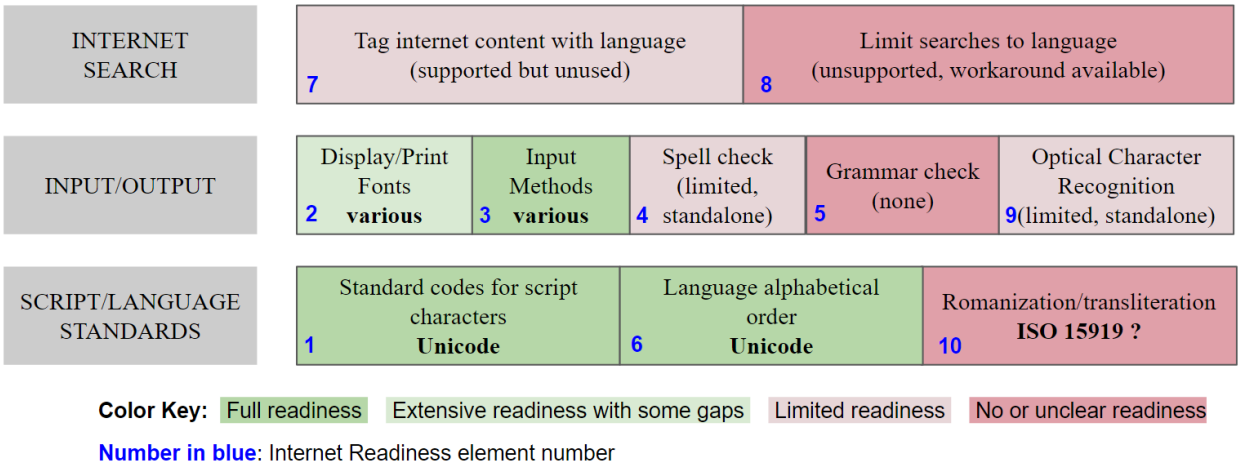

Figure 1: Internet readiness of Assamese in the technology stack

Here is a summary of the issues identified earlier in this paper where further work is needed for Assamese in the technology stack shown in Figure 1:

1. Integrating existing Assamese spell checkers into text editors
2. Developing grammar checkers for Assamese input
3. Tagging Assamese content on the internet with the language code 'as'
4. Getting Google to add Assamese as a supported language on Advanced Search
5. Developing full-featured Optical Character Recognition software for Assamese
6. Designing new fonts for Assamese

Another fundamental issue identified during the analysis is that Assamese text currently gets detected as Bengali language text on internet platforms like Google. The root cause is that neither content creators nor internet platforms are using existing standard mechanisms when it comes to distinguishing between Assamese and Bengali language texts. Internet platforms are making an assumption that all text using character codes from the Unicode code chart named Bengali must be in the Bengali language. This is although the specification for the Unicode code chart named Bengali is clear that it is shared between the Assamese and Bengali languages[1]. That assumption is a mistake and a violation of underlying standard mechanisms that exist in the technology stack to distinguish between languages that use a common code chart. Content creators also need to play their part by tagging their content with standard language codes to identify the language. These mechanisms are being used by content creators and the internet platforms to distinguish between English and Dutch content, for example. But these standard mechanisms are not being used to distinguish between Assamese and Bengali content yet. This needs attention from both content creators and from the internet platforms. Details of the issue in the Google platform and suggested fixes have been identified for communication to Google[27].

These are key priorities to increase internet readiness of Assamese and they need follow-up action. In addition to volunteer initiatives, work to focus on these priorities can be a source of income and entrepreneurial opportunity for talented software developers, graphic designers, and businesses. Focus and investment in these areas will increase internet readiness of Assamese. This will help everyone who wants to contribute to increasing the level of Assamese content on the Internet.

Further work

Beyond the ten essential elements analyzed here for internet readiness of a human language, additional elements need to be considered going forward. Some of those additional elements are:

1. Automated translation of text from one language to another
2. Natural speech synthesis of text in the language
3. Speech to text recognition for the language

As these and other new elements develop for Indic languages including Assamese, this comparative analysis will need to be extended to include those elements. Among these emerging topics, there is already keen interest in automated translation of text from languages like English to Assamese and vice versa. The recent launch of Assamese translation support in its products by Microsoft is a welcome initiative in this regard[28].

References

[1] "South and Central Asia-I: Official Scripts of India", Unicode® 13.0.0," *[Unicode]*, section 12.2, page 473. [Online]. Available: http://www.unicode.org/versions/Unicode13.0.0/ch12.pdf.

[2] "OpenType," *Wikipedia*, 11-Jun-2020. [Online]. Available: https://en.wikipedia.org/wiki/OpenType.

[3] লিটিকাই ৩.০ *(বানান পৰীক্ষক)*. [Online]. Available: http://www.xobdo.org/litikai.

[4] *LuitPad*. [Online]. Available: http://www.luitpad.in/.

[5] *Locale Data Summary for Assamese [as]*. [Online]. Available: https://unicode-org.github.io/cldr-staging/charts/37/summary/as.html.

[6] *Declaring language in HTML*. [Online]. Available: https://www.w3.org/International/questions/qa-html-language-declarations.

[7] *Google Advanced Search*. [Online]. Available: https://www.google.com/advanced_search.

[8] "Finding Assamese content on the internet: Can search engines eliminate the haystack?," Satyajit Nath. *Google Docs*. [Online]. Available: https://tinyurl.com/y8wr65ms.

[9] “Change your language on Google,” *Google Search Help*. [Online]. Available: https://support.google.com/websearch/answer/3333234.

[10] “Home,” *Pramukh OCR*, 17-Feb-2020. [Online]. Available: http://www.pramukhocr.com/.

[11] “Assamese transliteration,” *Wiktionary*. [Online]. Available: https://en.wiktionary.org/wiki/Wiktionary:Assamese_transliteration.

[12] “List of ISO romanizations,” *Wikipedia*, 26-May-2020. [Online]. Available: https://en.wikipedia.org/wiki/List_of_ISO_romanizations.

[13] “ISO 15919,” *Wikipedia*, 25-Apr-2020. [Online]. Available: https://en.wikipedia.org/wiki/ISO_15919.

[14] *Assamese Phonetic Keyboard for Mac OS*, www.xobdo.org/article/mac.

[15] *Assamese Phonetic Keyboard for Windows 7 & Windows 8*, www.xobdo.org/article/win7.

[16] “Lachit Multilingual Keyboard - Apps on Google Play.” *Google*, play.google.com/store/apps/details?id=com.lachit.android.assamese.

[17] “Google Indic Keyboard - Apps on Google Play.” *Google*, play.google.com/store/apps/details?id=com.google.android.apps.inputmethod.hindi.

[18] “Gboard - the Google Keyboard - Apps on Google Play.” *Google*, play.google.com/store/apps/details?id=com.google.android.inputmethod.latin.

[19] “Type in 23 Indian Languages.” *Pramukh IME*, 25 Apr. 2019, www.pramukhime.com/type.

[20] “Rodali Assamese Keyboard (ৰ'দালি) - Apps on Google Play.” *Google*, play.google.com/store/apps/details?id=com.sltdassam.rodali.

[21] *Locale Data Summary for Dutch [nl]*. [Online]. Available: https://unicode-org.github.io/cldr-staging/charts/37/summary/nl.html.

[22] *Locale Data Summary for Swedish [sv]*. [Online]. Available: https://unicode-org.github.io/cldr-staging/charts/37/summary/sv.html.

[23] “Europe-I: Modern and Liturgical Scripts”, Unicode® 13.0.0,” *[Unicode]*, section 7.1, page 287. [Online]. Available: http://www.unicode.org/versions/Unicode13.0.0/ch07.pdf.

[24] “South and Central Asia-I: Official Scripts of India”, Unicode® 13.0.0,” *[Unicode]*, section 12.2, page 447. [Online]. Available: http://www.unicode.org/versions/Unicode13.0.0/ch12.pdf.

[25] “Vidura—an interactive multilingual publishing system—specification & design,” Satyajit Nath, Sumanta N. Pattanaik, S.P. Mudur. Proceedings of the International Conference on Electronic Publishing on Document manipulation and typography, May 1988. Pages

249–260. https://dl.acm.org/doi/abs/10.5555/51292.51311.

[26] “Lohit Assamese Font by Lohit Fonts Project Free Download,” Lohit Fonts Project. [Online]. Available: https://www.fontsc.com/font/lohit-assamese.

[27] “Assamese in Google products: Usability Challenges," Satyajit Nath. Private document as input for filing issues on Google product forums, July 2020.

[28] “Microsoft Translator adds Assamese, strengthens support for 12 Indian languages," Microsoft. [Online]. Available: https://news.microsoft.com/en-in/microsoft-translator-adds-assamese-strengthens-support-for-12-indian-languages/.