On the Characteristics of Language Tags on the Web

Joel Sommers

Colgate University, Hamilton NY, USA
jsommers@colgate.edu

Abstract. The Internet is a global phenomenon. To support broad use of Internet applications such as the World Wide Web, character encodings have been developed for many scripts of the world's languages and there are standard mechanisms for indicating that content is in a particular language and/or tailored to a particular region. In this paper we study the empirical characteristics of *language tags* used in HTTP transactions and in web pages to indicate the language of the content and possibly the script, region, and other information. To support our analysis, we develop a new algorithm to infer the value of a missing language tag for elements used to link to alternative language content. We analyze the top-level page for websites in the Alexa Top 1 Million, from six geographic perspectives. We find that one third of all pages do not include any language tags, that half of the remaining sites are tagged with English (`en`), and that about 10K sites have malformed tags. We observe that 80K sites are multilingual, and that there are hundreds of sites that offer content in the tens of languages. Besides malformed tags, we find numerous instances of correctly formed but likely erroneous language tags by using a Naïve Bayes-based language detection library and comparing its output with a given page's language tag(s). Lastly, we comment on differences in language tags observed for the same site but from different geographic vantage points or by using different client language preferences via the HTTP `Accept-Language` header.

1 Introduction

The Internet and World-Wide Web were originally designed by a relatively homogeneous, English-speaking group of engineers and scientists with no explicit technical concern for supporting languages other than English [3]. Although early web designs used ASCII character encoding and lacked any provision for indicating the language of the text within an HTML page, both HTTP and HTML have evolved to support character encodings for many scripts of the world's languages, and to support multiple ways for indicating the language of a page or elements within a page [2,4]. The culmination of these capabilities is that today, web browsers routinely inform web servers about a user's language preferences through the HTTP `Accept-Language` header [4], servers can use the expressed preferences to deliver desired content, if it is available, and browsers can display text in the native script of a user's preferred language.

Language tags are used to indicate the language(s) preferred by a client, or the language of text or elements within content delivered by a server[1]. Language tags provide important context for web content and there are a number of reasons why it is critical that they be constructed correctly and used in ways to enhance the semantics of web pages and elements within pages. First, browsers may use language tags for rendering content, *e.g.*, right-to-left rendering for some languages, or highlighting/translating content in a user's preferred language. Second, appropriate language tags on internationalized and localized pages can help search engines respond to queries with appropriate content and thus increase site traffic and ad revenue[2]. Third, screen readers for the visually impaired may use language tags for determining whether to read content within a page or whether to ignore content. Fourth, language tag attributes on hyperlinks (*e.g.*, `hreflang` tag within `<a> elements`) can be used to indicate the availability of alternative language content; browsers may use such attributes to help users find preferred content. Lastly, including appropriate language tags can help speakers of underserved (so-called "minority") languages to find and better utilize content. Understanding the nature of how language tags are used across the web may provide a useful perspective on how to improve access to desired content and to bridge the global digital divide.

In this paper we analyze the empirical characteristics of language tags found in HTTP response headers and within HTML pages. We gather data from the Alexa Top 1 Million sites from six geographic vantage points by using a commercial VPN service. We focus on the top-level document (URI path /) for each site, recognizing that this may not give a comprehensive view of a site's language offerings. We perform two types of requests: one in which the `Accept-Language` (`A-L`) header is set to `*` to accept *any* language and one in which the `Accept-Language` header is set to a list of (*de jure* or *de facto*) official languages or commonly-used languages within the same region from which we launch our requests. We refer to the data collected using `Accept-Language: *` as our *default* language data and to the data collected using a region-specific `A-L` header as *langpref* data. We collect both HTTP response headers and the content of the top-level document; we do not access any linked resources (*e.g.*, JavaScript, iframes, images, etc.) nor do we execute any JavaScript code. In total, we performed 12 million web requests (not including retries due to transient errors), collecting a total of about 500 GB of compressed HTTP headers and content for analysis.

For each `A-L` variant (default and langpref) and for each VPN location, we extract language tags that indicate the *primary* language of content on the page, and we also extract language tags from every element within the page in order to gain a perspective on the breadth of languages in which content is offered for a given site. Specifically for hyperlink elements, we describe an algorithm for inferring the value of a missing language tag for links that are used to lead

[1] The structure and valid values of language tags are specified in IETF BCP 47 [9] and the IANA language subtag registry [1], respectively, as discussed below.

[2] `https://searchenginewatch.com/sew/howto/238631/localization-for-international-search-engine-optimization`

to alternative language content on the same site. We find that, overall, one third of all pages do not include any language tags, and that another third of pages are tagged with English as the primary language. We find that our inference algorithm contributes to 1–3% of language tags found, which varies depending on VPN vantage point and the default and langpref `A-L` header. We observe that 80K sites out of the Alexa Top 1 Million are multilingual, and that about 30K of those sites offer content in two languages with some of the remaining sites offering many tens of languages. We find that nearly 1% (about 10K sites) of all language tags are malformed, and we find additional instances of correctly formed but likely erroneous language tags by using an off-the-shelf Naïve Bayes-based language detection library and comparing its output with a page's primary language tag. Lastly, we comment on region, script, and private-use subtags observed within language tags, and differences observed across VPN vantage points. We note that the code used to perform our study is publicly available [3] and our data will be made publicly available.

2 Background and Related Work

The structure and content of language tags used by Internet protocols and applications is described in IETF BCP 47 [9]. Language tags are formed from one or more *subtags*, which may refer to a language, script, region, or some other identifying category. The simplest language tag can include just a language subtag (*e.g.*, `en` (English), `de` (German), `cy` (Welsh)), but BCP 47 permits script subtags (*e.g.*, `Cyrl` (Cyrillic)), region subtags (*e.g.*, `AR` (Argentina)), and private-use subtags, among other features [7]. In practice, it is common for language tags to include between one and three subtags, *e.g.*, `es` (Spanish, not specific to any region), `pt-BR` (Brazilian Portuguese), `zh-Hant-CN` (Chinese, Traditional script, in China). Valid subtags within the categories defined in BCP 47 are detailed in the IANA language subtag registry [1], which serves as a kind of meta-registry of tags defined by other standards organizations.

The choice of a language tag to use in relation to web content may not be simple, and the W3C offers guidance on forming a language tag (keep it as short as possible) and how to correctly use tags in HTML documents [7,10]. The latest guidance regarding HTML is that the language tag for a page should be specified in the `lang` attribute of the top-level `<html>` element. If any divisions within a page are targeted at speakers of different languages, each of those divisions should similarly include an appropriate `lang` attribute. For links on a page that lead to alternative language content, the `hreflang` attribute can include an appropriate language tag [8], or `<link rel=alternate>` tags can include a URI to an alternative representation (*e.g.*, different language content) [4].

Unfortunately, previous versions of HTML and XHTML have used different mechanisms for indicating the language of a page and of elements within a page. For example, XHTML defines an attribute `xml:lang` which plays a similar role as the `lang` attribute in HTML5 [8]. Moreover, the HTTP response header

[3] `https://github.com/jsommers/weblingo`

[4] `https://searchenginewatch.com/sew/how-to/2232347/a-simple-guide-to-using-rel-alternate-hreflang-x`

`Content-language` has also been defined to indicate the intended language audience of a response, and particular `<meta http-equiv=content-language>` tags have been used to convey the same information. When multiple language indications are present on a page, the guidance provided by W3C from a *browser* perspective to determine the primary language of a server response is to first prefer the `lang` or `xml:lang` attributes if they are present, followed by the `<meta>` header if present, followed by the HTTP `Content-Language` header if present.

Web browsers may also inform servers of a user's language preferences through the HTTP `Accept-Language` header [4,5]. The value supplied in this header can be one or more language tags, with optional *quality* values indicating an order of preference. Quality values range from 1 (most preferred) to 0 (not wanted). For example, `cy;q=0.9, en;q=0.5, *;q=0.3` indicates that Welsh (`cy`) is most preferred, following by English, followed by anything else. The process of a server matching content preferences indicated by HTTP `Accept-` headers and available resources is known as *content negotiation* [5]. Although it can be unclear from a client's perspective how a server has decided to return a specific version of a resource, HTTP servers *should* include a `Vary` header indicating the parts of a request that influenced a server's decision.

There has been little prior work on studying language tags within HTML pages and in HTTP transactions. One recent work is [11], in which the authors report on the top 10 most common language tags found in `A-L` headers from clients that were making a request for a JavaScript instrumentation library. Besides that paper, the most closely related efforts are works that have sought to survey the number of documents available in various languages on the web. In [6] and references therein, the authors state that as of 1997, English was the language of 82.3% of pages, "followed by German (4.0%), Japanese (3.1%), French (1.8%) and Spanish (1.1%)". The dominance of English was also observed in [12] in 2002 (68% of pages), with increases in Japanese and Chinese content. We are not aware of studies that have focused specifically on evaluating language tags available in HTML pages and in HTTP transactions.

3 Methodology

To drive our empirical analysis of language tags we developed a web crawler in Python, leveraging the widely-used `requests` module, along with the `certifi` module to enable better TLS certificate verification[5]. We set the `User-Agent` string to a value equivalent to a recent version of the Google Chrome browser, and configured `requests` to allow up to 30 redirects before declaring failure. We also set connection and response timeouts to conservative values of 60 seconds each. We configured our crawler host to use the Google public DNS servers (8.8.8.8 and 8.8.4.4) and parallelized our crawler to speed the measurement process.

We used the Alexa Top 1 million sites as the basis for our study[6]. Although this list of websites is crafted from the point of view of one (albeit very large) cloud provider, we argue that it is adequate for gaining a broad view of today's

[5] `http://docs.python-requests.org/en/master/`

[6] `http://s3.amazonaws.com/alexa-static/top-1m.csv.zip`

web. In our future work we are considering how to expand the scope of the websites under study in order to measure a larger portion of the web and to improve coverage of sites that serve content for less dominant languages. For each web site, we made a request for the top-level resource (URI path `/`). Although accessing one document on each site may not give a comprehensive picture of a site's possible multilingual offerings, we argue that since it is anecdotally commonplace for a provider to link to different versions of a site from the top-level URI, it should still give a reasonably complete view.

It is also common for different sites to geolocate clients in an attempt to deliver appropriate content. To account for this, we used a commercial VPN service and launched requests through six different geographic locations. Moreover, for each site, we made two requests using two different versions of the HTTP `Accept-Language` header. In the first, we set the header to accept any language (`*`), and in the second we set the header to include a prioritized set of languages based on the *de jure* or *de facto* official languages of the VPN location used; we use the curated GeoNames.org list of country codes and languages for this purpose[7]. Table 1 lists the specific country codes and language preferences for the six VPN locations we used. For each of these language preferences, we explicitly set English to be least preferred given its traditional dominance in web content.

For each request and response exchange, we store the full HTTP request and response headers, along with the full (compressed) response content and metadata such as the time a request started and ended and the original hostname used in the request. We retried any errored requests up to three times, storing error information in our logs, as well as any information about redirects. No additional requests were made for directly linked content, such as JavaScript, CSS, image files, or iframes. We did not execute any embedded JavaScript. We note that anecdotally, some sites use JavaScript to dynamically add widgets to allow a user to select a preferred language. Due to our measurement methodology, we missed any of these instances that would have included explicit or inferable language tags. In our future work we intend to quantify the number of sites that use such techniques.

Overall, we made 12 million web requests, not including retries because of transient errors (1 million sites, 6 VPN locations, 2 `A-L` header values), resulting in approximately 500 GB of request and response data. For each instance of VPN location and `A-L` header value, there were approximately 70K requests that resulted in unrecoverable errors. The most common error was DNS failure ($\approx 50K$) followed by connection failures and timeouts ($\approx 19K$). We also observed TLS errors ($\approx$ 200), content decompression errors ($\approx$ 500), and a handful of internationalized domain name errors ($\approx$ 15).

We used the Python `BeautifulSoup4` module[8] with the `lxml`[9] parser to analyze content and extract language tags. There was no existing Python module to rigorously validate language tags for structure and content, so we created one

[7] `http://download.geonames.org/export/dump/countryInfo.txt`

[8] `https://www.crummy.com/software/BeautifulSoup/`

[9] `http://lxml.de`

as part of our work[10]. Our module conforms to BCP 47 and enables validation and extraction of subtags within a language tag. We also used the `langcodes` module for analyzing text on pages (used in our inference algorithm, described below)[11]. While this module can also parse language tags, we found it to accept tags that would not be considered valid by BCP 47. Lastly, we used a Python port of the Compact Language Detector (`pycld2`[12]) to detect the language within text on pages. Internally, this module uses a Naïve Bayes classifier to detect the language. It is widely used and includes support for 165 languages.

Table 1. Accept-Language headers used for langpref (non-default language) experiments. The region code refers to the country from which HTTP requests are launched.

Region	`Accept-Language` value in HTTP requests
AR	`es-AR;q=1.0, es-419;q=0.9, es;q=0.7, it;q=0.6, de;q=0.4, fr;q=0.3, gn;q=0.1`
GB	`cy-GB;q=1.0, cy;q=0.8, gd;q=0.6, en-GB;q=0.4, en;q=0.2`
JP	`ja;q=1.0`
KE	`sw-KE;q=1.0, sw;q=0.8, en-KE;q=0.5, en;q=0.2`
TH	`th;q=1.0`
US	`es-US;q=1.0, es;q=0.8, haw;q=0.7, fr;q=0.5, en-US;q=0.3, us;q=0.2`

We observed in our initial analysis that there were many sites that did not include `lang` or `hreflang` attributes (or any other metadata) to indicate that a hyperlink leads to alternative language content. As a result, we developed an algorithm for analyzing hyperlink (`<a>`) tags to determine whether a language tag should be *inferred*. The basic approach of our algorithm is to extract several components from a link tag: (1) the domain (if any) in the `href` attribute, (2) the URI path in the `href` attribute, (3) any query parameters in the URI, (4) keys and values for other attributes within the tag, and (5) the text. We analyze each of these components for "language indicators":

- For the domain, we match the left-most domain (most specific) with language and/or country subtags. For example, `https://es.wikipedia.org` contains Spanish-language content.
- For the URI path, it is not uncommon for web sites to include the language tag (and possibly region subtag) in the first one or two components. For example, `http://www.ikea.com/us/en/` provides English content for US-based users.
- For some sites, query parameters are used to indicate the language. For example, Google uses the query key `hl` (for "human language") to indicate the language, such as `https://www.google.com/?hl=cy`.

[10] `https://github.com/jsommers/langtags`
[11] `https://github.com/LuminosoInsight/langcodes`
[12] `https://github.com/aboSamoor/pycld2`

– Other sites use non-standard attributes (*i.e.*, not `lang` or `hreflang`) to indicate the language of the linked content. We've observed sites to use `data-lang=de` as an attribute to refer to German content, for example.
– Lastly, the clickable text is often either a language subtag or the name of a language, *in the language of the page content.* We use the `langcodes` module to map language names to subtags.

The algorithm seeks to match the inferred language tag from at least two indicators, and requires that the original text harvested from two indicators be *different.* This requirement is to avoid false inferences in situations such as when the subdomain matches a valid language tag (*e.g.*, `ru`) and the link text *includes* the word Russia, but also includes other words (*e.g.*, "News from Russia"). Table 2 shows examples of what our algorithm would infer for three (real) example links. Through extensive manual inspection of links and inferences, we found that our algorithm is conservative in the sense that it does not make inferences on *all* links that lead to alternative language content, but the inferences it makes are sound. In other words, in our manual inspections we observed some false negatives, but no false positives. In future work we plan to examine how our algorithm can be improved. Overall, we found that our inference method contributed about 1–3% of all language tags found. Of all the tags found through the inference algorithm, approximately 5% were tags that had not been previously observed, *i.e.*, the number of total language tags observed was expanded via our inference method.

Table 2. Examples of hyperlinks links from which the language tag can be inferred.

Markup	**Language inferred**
`<a href="#" rel="lt">litvn</a>`	Lithuanian (lt) (Hungarian site)
`<a class="site-topbar__link" href="/en/personal" tabindex="0"><span class="site-topbar__langs__text"> English</span></a>`	English (en) (Swedish site)
`<a href="javascript:setLang(’ar’);"> Arabic</a>`	Arabic (ar) (English site)

4 Results

In this section we describe the results of our analysis. We begin by discussing the prevalence of malformed language tags and sites that do not include any language tags. As noted above, various types of errors prevented data collection for approximately 70K of the 1M sites. For the remaining sites, about 330K do not include any language tags at all; this number varies between 329K and 335K depending on `A-L` setting and VPN location. Of sites that include language tags, we observed about 4K sites to use malformed *primary* language tags (across all language tags, not just primary, about 10K were malformed). To extract the primary language tag for a page, we first consider the `lang` (or xml:lang) attribute, followed by any <meta> header, followed by any HTTP Content-Language header, in that order. Table 3 summarizes the most common types of errors we found. Other malformed tags included HTML fragments, apparent Boolean values (*e.g.*, False), apparent "codes", and other garbage. Considering all the malformed tags, it is clear that they fall into one of two categories: semantic errors (*e.g.*, including a region subtag instead of a language tag) or programmer/developer errors (*e.g.*, uninterpolated language variables).

Next, we examine the collection of valid primary language tags found for each site from various geographic perspectives. Figure 1 shows a bargraph for the top 30 most frequent language tags found, from each VPN vantage point. Data are shown for the default A-L header. We note that the data shown comprise about 85% of all valid primary language tags and that the tail is long: there are around 180 distinct primary language tags discovered. As for *how* the primary language tags were found, on average about 94% come from the <html> tag's lang attribute, another 2% come from the xml:lang attribute, 1.5% come from the <meta http-equiv ...> header, and 2.5% come from the HTTP Content-language header.

We observe in the figure some apparent effects that IP geolocation has on the language tag presented in the response. For example, with the TH vantage point we observe an increase in occurrences of the th language code. There are similar increases for es in AR and ja in JP (which are relatively smaller due to the log scale) and for sw in KE (not shown in the plot).

We observe two clear apparent anomalies in the plot: the dip in ko for GB, and the dip in fa for JP. Regarding the Korean language tag, we observe similar patterns for the ko language tag for other vantage points with a non-default A-L header (shown below). It appears that there are a large number of Korean sites that erroneously include the ko subtag but for non-Korean language content. It is unclear presently which set of parameters causes a change in these sites' behaviors, but it may be that a commonly used library or service within Korea is at the root of the issue. For the fa dip observable from the JP vantage point, we are not yet able to speculate on the cause. Interestingly, we also observe a slight rise in uk (Ukrainian) for GB, which is likely due to misusing a region tag (which should be GB in any case).

Table 3. Most common malformed tag types.

Type of problem	**Example**	**% of total**
Country/region code used as language code	cn	32%
Language name	Deutsch	17%
Character encoding instead of language tag	UTF-8	5%
Non-interpolated placeholder	{{ currentLanguage }}	4%
Other malformed tags		42%

We also examined differences in the primary language tags observed between the default A-L and langpref A-L for each vantage point. For the TH vantage

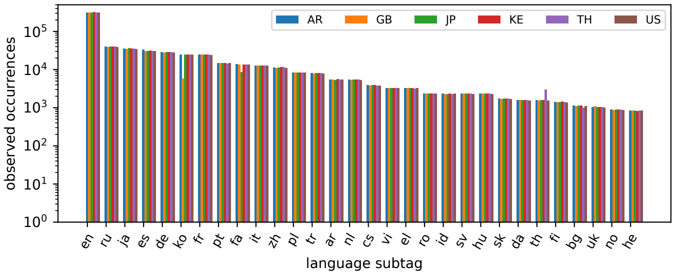

Fig. 1. Primary language observed for all vantage points for top 30 most frequently observed languages, with default `Accept-Language` header. Note the log scale.

point (not shown due to space constraints), we observe an increase in occurrences of `th` as the primary language when the `A-L` header is set to strongly prefer Thai language content. Beyond that, however, the impact of content negotiation due to the non-default (langpref) `A-L` header is unclear. Specifically, we observe increases in non-Thai language subtags (*e.g.* in `bg` (Bulgarian)!), and similar phenomena are observed in data collected from other vantage points. In particular, compared with accepting a default language, the specific `A-L` header causes an *increase* in the occurrences of primary language tags for languages that are not even included in the preference list. Further, we note that the HTTP specification states that servers *should* include a `Vary` header indicating which client preferences went into determining the content delivered [5]. However, we observe a mere 40 sites that include an indication of `Accept-Language` in the `Vary` header response, which is far below the number of (fairly significant) differences we observe. Clearly, content negotiation plays a larger role than is indicated by the `Vary` header, which we intend to investigate as part of our future work.

In Figure 2 we show the distribution of the total number of language subtags observed across all sites. From this figure we see that about 330K contain zero language subtags (far left bar), and that about 520K sites are apparently unilingual (*i.e.*, we observe a single language subtag). On about 80K sites, we observe more than one language subtag. From this, we infer that about 80K sites of the Alexa 1M are multilingual; of those, about 30K are bilingual. At maximum, we observe 376 distinct language subtags on one site (not shown in the figure) and at least 45 sites offer some content in 100 languages or more. From our analysis, it is not clear *how much* content is offered in any given language, although we believe that the fact that *any* content is offered multiple languages to be of interest for the purpose of our study.

Next, we aggregate the counts of all language subtags discovered across all sites and show the distribution in Figure 3, showing the top 50. We cannot view this distribution as giving an accurate sense for the prevalence of various languages across the web, but we believe that the figure nonetheless provides an interesting view of language diversity on the web. Of note in the figure are the large number of occurrences the private language tag `x-default`, which the

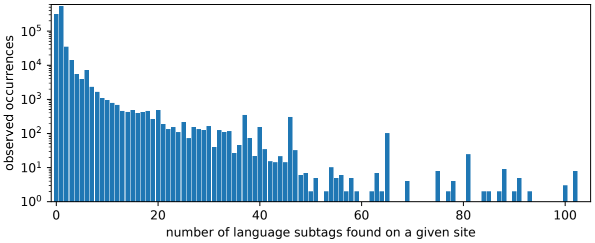

Fig. 2. Number of languages observed to be offered per site. Note the log scale.

W3C recommends to avoid whenever possible [10] (although we note that very few of these appear as the primary language tag). Also, we observe the presence of two languages that are on the UNESCO endangered languages list[13] (each with a status of Vulnerable): Belarussian (`be`) and Basque (`eu`). Lastly, we note that our ranking of most prevalent languages observed on the web differs from those published in prior work [6,12]. In particular, Russian and Japanese appear much more frequently than when those prior studies were done.

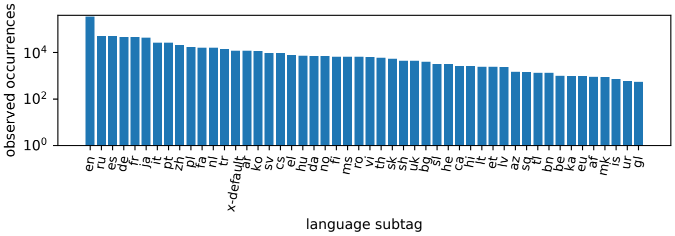

Fig. 3. Frequency of language tags seen across all sites. Note the log scale.

Next, we consider additional components in language tags, such as the region subtag. In Figure 4, we show the distribution of region subtags observed in primary language tags from the KE vantage point (langpref `A-L` setting). First, we note that a total of 284K sites included a region subtag (which is a fairly consistent figure across all vantage points and `A-L` settings), and we observed a total of 227 distinct region subtags. For some vantage points, we observe many fewer distinct region subtags (as few as 160). Interestingly, we observe that the `KE` region subtag is the second most common. We infer from this (and results from other vantage points) that many sites geolocate client IP addresses and blindly include a region code based on client location. We note also that the inclusion of the region subtag runs counter to W3C advice [10], which is to only include the region subtag when it provides distinct information about site localization.

[13] `http://www.unesco.org/languages-atlas/`

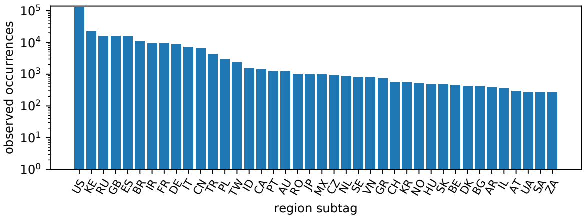

Fig. 4. Frequency of observed region subtags in primary language tags observed in data collected from the KE vantage point. Note the log scale.

Lastly, we compare primary language subtags with the result of using the `pycld2` language detection library on the content. For this analysis, we only consider sites/pages for which a primary language tag is included and valid since we wish to understand whether a given language tag is likely to be accurate. Figure 5 shows results for the default `A-L` setting for the AR vantage point. The top 40 most frequently occuring primary language subtags are shown. We observe in the figure that in many cases, the primary language subtag is close-to-correct. Interestingly, it appears that there are a number of pages in Hindi (`hi`) that are mis-tagged (though it may also be that `pycld2` is incorrect for some of these cases). Information from this analysis could be used to inform sites of misconfigurations, or suggestions to improve the localization of their site by including the appropriate language tag(s).

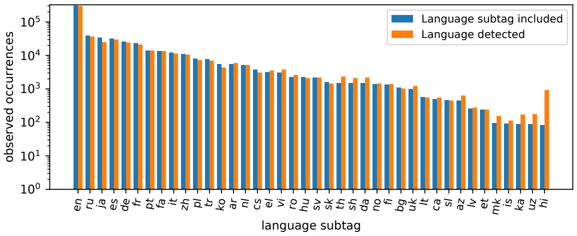

Fig. 5. Primary language tag versus language detected using a Naïve Bayes language detection library, for the AR vantage point. Note the log scale.

5 Summary and Conclusions

In this paper we study the empirical characteristics of language tags observed on web pages and in HTTP transactions. We examine the sites in the Alexa top 1 million, gathering data from six geographic vantage points and using two different settings for the HTTP `Accept-Language` header. We find that about 1/3 of all sites do not include a primary language tag, and that English (`en`) is the most commonly occurring language subtag. We find many occurrences of malformed tags and that about 8% of sites are multilingual. We analyze the prevalence of different language subtags across all sites and vantage points, and comment on various anomalies observed in the data.

In our ongoing work, we are considering a number of directions. First, we plan to examine how HTTP content negotiation affects language tag inclusion, and how it may impact users who are trying to find content in their preferred language. We are also examining ways to broaden our language subtag inference algorithm to consider other elements (*e.g.*, form entries) that indicate the availability of alternative language content on a site. Lastly, we are looking at ways to expand our study beyond the Alexa top 1 million list in order to gain a more comprehensive view of human language on the web.

Acknowledgments

We thank Alex Nie '20 and Ryan Rios '20, who contributed to earlier stages of this work for their summer research. We also thank Ram Durairajan for insightful comments on this work, as well as the anonymous reviewers. Lastly, we thank the Colgate Research Council, which provided partial support for this research.

References

1. IANA Language Subtag Registry. `https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry`
2. World Wide Web Consortium. Internationalization techniques: Authoring HTML and CSS. `https://www.w3.org/International/techniques/authoring-html` (January 2016)
3. Abbate, J.: Inventing the Internet. MIT Press (2000)
4. Fielding, R., Reschke, J.: RFC 7230: Hypertext Transfer Protocol (HTTP/1.1): Message Syntax and Routing. `https://tools.ietf.org/html/rfc7230` (June 2014)
5. Fielding, R., Reschke, J.: RFC 7231: Hypertext Transfer Protocol (HTTP/1.1): Semantics and Content. `https://tools.ietf.org/html/rfc7231` (June 2014)
6. Grefenstette, G., Nioche, J.: Estimation of English and non-English language use on the WWW. In: Content-Based Multimedia Information Access-Volume 1. pp. 237–246 (2000)
7. Ishida, R.: Language tags in HTML and XML. `https://www.w3.org/International/articles/language-tags/`
8. Ishida, R.: Declaring language in HTML. `https://www.w3.org/International/questions/qa-html-language-declarations` (2014)
9. Phillips, A., Davis, M.: Tags for Identifying Language. `https://www.rfc-editor.org/rfc/bcp/bcp47.txt` (September 2009)
10. R. Ishida: Choosing a Language Tag. `https://www.w3.org/International/questions/qa-choosing-language-tags` (2016)
11. Thomas, C., Kline, J., Barford, P.: IntegraTag: A Framework for High-Fidelity Web Client Measurement. In: Teletraffic Congress (ITC 28), 2016 28th International. vol. 1, pp. 278–285 (2016)
12. Xu, F.: Multilingual WWW. Knowledge-based information retrieval and filtering from the web 746, 165 (2003)