Comparing the Usage of Non-Latin Character Sets in Top-Level Domains

Haylin Moore
University of Massachusetts Amherst
haylinmoore@umass.edu

***Abstract*—The domain name system, which translates domain names to IP addresses, only allows for alphanumeric characters and hyphens. Meaning that languages not based on Latin can not have domains using their language's native characters. Punycode is a technology that encodes any Unicode character into the domain character set. It is seen that punycode is being mainly used by for Chinese, Japanese and Korean (CJK) text, making up 57 percent of Punycode based top-level domains (TLDs). CJK domains made up the top seven Punycode TLDs, and 96 percent of CJK second-level domains (SLDs) also use a CJK TLD.**

***Index Terms*—domain name system, internet**

I. Introduction

The Domain Name System (DNS) is a critical service required for the modern web that translates domain names to IP addresses and vice versa. DNS is limited to only the alphanumeric character set and hyphens, with the form for a domain defined as <let-or-digit>[*[<let-or-digit-or-hyphen>]<let-or-digit>] [1], [2]. Previously there was no way for a domain to use non-alphanumeric characters until the introduction of Punycode in 2003 [3]. Punycode is simple, efficient, and fully supports all arbitrary code points [3]. Punycode domains start with "xn–" then any alphanumeric characters, followed by a hyphen with more alphanumeric characters that encode any non-alphanumeric characters. For example, "exámple" would become "xn–exmple-qta" domains without any alphanumeric characters just immediately go into the encoding like so "xn–ses554g" [3]. Browsers will automatically convert to and from the "xn–" form of Punycode domains to the Unicode version when displaying it to the user.

II. Data Collection

A. Centralized Zone Data Service

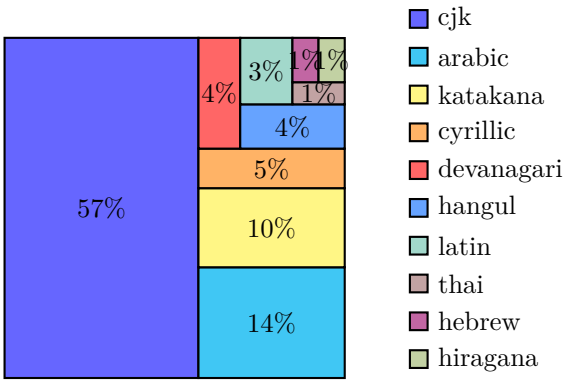
The Internet Corporation for Assigned Names and Numbers (ICANN) is the organization in charge of creating top-level domains (TLDs). A TLD is the ending chunk of a typical website domain for instance the "com" is the TLD for "google.com". They also run a service called the Centralized Zone Data Service (CZDS) which acts as a portal for individuals to request access to the Zone Files for participating TLDs. A zone file is a text file that holds the Resource Records for a domain, for instance, the com zone file holds a list of all com domains [4]. Through a CZDS request, the researcher got access to 79 different zone files for Punycode TLDs.

B. Character Set Identification

From the 79 zone files, a total of 492,000 second-level domains (SLDs) were discovered. These SLDs were then scanned for their character sets using a modified tool derived from software by BabelStone which identifies the Unicode character classification for sets of characters. For each SLD, the character set for all characters in the SLD was found, then the most common character set, excluding Latin, was used as the SLD's character set. The same process was also done on the TLDs.

III. Analysis of Data

A. Character set breakdown of TLDs



B. Most Popular Punycode TLDs

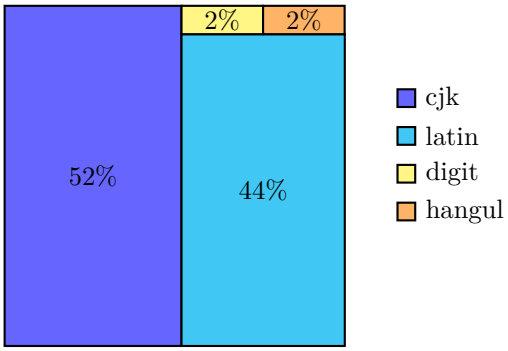
The top seven TLDs were all Chinese, Japanese, and Korean (CJK), with CJK making up 11 of the top 15 punycode TLDs. Due to sharing characters Unicode has classified them as one Unicode block and for this paper was treated like so. Interestingly despite Arabic TLDs making up 14 percent of all punycode TLDs and being the second most common character set for punycode TLDs, there is not a single Arabic TLD until the 24th spot, when ranked by usage, with it only have 902 SLDs registered on it. This also shows how infrequently punycode TLDs are actually used with the 15th most used punycode TLD "xn–q9jyb4c" not even breaking the mark of having 2,000 SLDs.

TABLE I
MOST POPULAR TLDS WITH THEIR CHARACTER SET AND COUNT

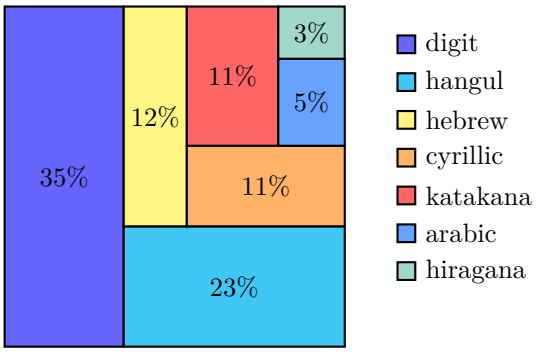
TLD	Character Set	Count
xn–ses554g	cjk	281,140
xn–czr694b	cjk	46,618
xn–czru2d	cjk	43,452
xn–55qx5d	cjk	33,021
xn–3ds443g	cjk	23,699
xn–io0a7i	cjk	22,355
xn–hxt814e	cjk	9,648
xn–mk1bu44c	hangul	4,773
xn–tckwe	katakana	3,336
xn–fiq228c5hs	cjk	2,691
xn–g2xx48c	cjk	2,306
xn–vuq861b	cjk	2,248
xn–vhquv	cjk	2,207
xn–9dbq2a	hebrew	2,110
xn–q9jyb4c	hiragana	1,881

C. Distribution of character sets in SLDs

Looking at the character sets used in the SLDs one sees an even more extreme version of the TLD chart. It shows the CJK character set being the overwhelming majority.

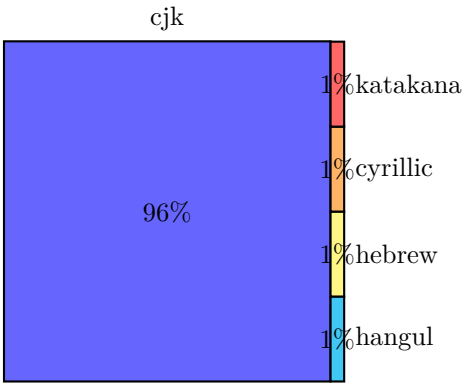


Below is a graph which removed by CJK and Latin character sets from the breakdown.

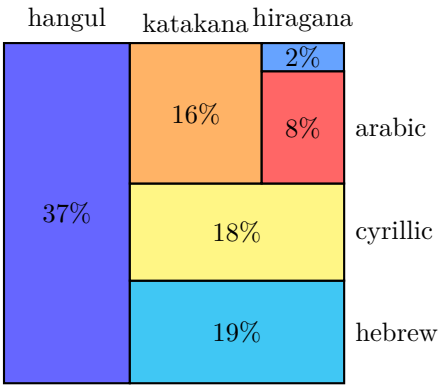


D. SLDs With The Same Character Set as Their TLDs

Looking at the character set used in SLDs and comparing them to their TLD's characters set, it is seen that once again CJK makes up the vast majority.



Below is the same graph but with the CJK character set removed so that the distribution for non-CJK character sets can be better visualized.



IV. CONCLUSION

Punycode adoption is the strongest in Asian countries, but even with the adoption of these TLDs their usage still pales in comparison to common TLDs like .com which has over 140 million domains registered to it [5].

ACKNOWLEDGMENT

Thank you to ICANN for creating the CZDS, which allowed for this research to happen, and each of the TLDs which graciously gave me access to their zone files.

REFERENCES

[1] M. S. K. Harrenstien and E. Feinler, "Dod internet host table specification," RFC 952, October 1985.
[2] R. Braden, "Requirements for internet hosts – application and support," RFC 1123, October 1989.
[3] A. Costello, "Punycode: A bootstring encoding of unicode for internationalized domain names in applications," RFC 3492, March 2003.
[4] M. Lottor, "Domain administrators operations guide," RFC 1033, November 1987.
[5] (2020, September) What does a registrar do? [Online]. Available: https://glauca.digital/blog/2020/07/30/what-does-a-registrar-do.html