Thesis

Gabe DeFreitas

October 21, 2018

Contents

1 Introduction **1**

2 Very Long Names **1**

3 Birth Certificates **2**
3.1 California 2
3.1.1 Proposition 63 2
3.1.2 2014-AB-2528 2
3.1.3 2017-AB-82 3
3.2 Massachusetts 4
3.3 New Hampshire 4
3.4 Kansas 4

4 Passports **4**
4.1 Latin 5
4.2 Cyrillic 5
4.3 Arabic 5

1 Introduction

In our "digital age" writing has transitioned from a free-flowing calligraphic artform to a discretized logical structure consisting of well-defined glyphs. The fact that most of the early development of computers took place in the United States gave English an ascendancy over other languages for the purposes of digital communication. English, perhaps as a coincidence, is also one of the easiest languages to represent in code, requiring at the bare minimum just the 26 non-accented characters of the English alphabet, and perhaps some punctuation and numbers mixed in.

2 Very Long Names

> You see, to some people in the world, your name is everything. If I say my name to an elder Hawaiian (kupuna), they know everything about my husband's family going back many generations... just from the name. When the name is sliced up, changed or altered it distorts the intention and meaning that the name represents. Unfortunately, many people have been shamed into hiding their real names because they don't fit in with the dominant culture's lack of respect for the name.
>
> P.S. If bills or traffic citations are not correctly addressed, my husband refuses to pay and is under no obligation legally to pay.

3 Birth Certificates

American law generally holds the naming of children to be the right and responsibility of parents, without shutting the door on regulating edge cases, like "Ghoul Nipple", "Legend Belch", "Brfxxccxxmnpccccllmmnprxvclmnckssqlbb11116", and "" [Larson11]. In many American states, however, this right is abridged with reference to diacritical marks above letters, hardly an edge case in many languages around the world. [Larson11] investigates this in his study of American naming law, finding states with such rules to include California, Massachusetts, New Hampshire, and Kansas:

3.1 California

Guidelines provided by the California Office of Vital Records instruct county clerks that baby names may contain only "the 26 alphabetical characters of the English language with appropriate punctuation if necessary" and that "no pictographs, ideograms, diacritical marks (including "é," "ñ," and "ç") are allowed" [Larson11].

The handbook cites Proposition 63, the 1986 ballot referendum in which Californian voters declared English the state's official language, as legal justification. Larson points out, however, that the names of two California state parks, Año Nuevo State Park and Montaña de Oro State Park, manage to contain such characters. Moreover, the City of San José, California includes the acute accented é in its official name, and its Style Guide even includes instructions on how to produce it electronically: "To create an accented é, hold down the alt key and type "0233", on the numeric key pad." California's Department of Public Health likely disobeys the city's guidelines in birth certificates, though this needs to be verified.

3.1.1 Proposition 63

3.1.2 2014-AB-2528

A 2014 bill in the California State Assembly sponsored by AM Nancy Skinner (AB-2528) sought to rectify the state's processing of birth certificates and driver's licenses by allowing diacritical marks in names. The bill "required the State Registrar to ensure that diacritical marks on English letters are properly recorded on birth certificates, death certificates, certificates of fetal death, and marriage licenses, including, but not limited to, accents, tildes, graves, umlauts, and cedillas" [AB-2528]. This bill stalled in the Appropriations Committee when state agencies predicted multi-million dollar price tags relating to IT upgrades, noting that the DMV's software could not "even accept lower-case letters". For this same reason the bill was opposed by the County Recorder's Association of California.

3.1.3 2017-AB-82

In 2017, California AM Jose Medina revived the issue with AB-82, which ultimately passed both houses of the legislature before being vetoed by Governor Jerry Brown. Unlike the 2014 bill, this edition did not affect the issuance of driver's licenses, only birth certificates. Passing through many more stages of the legislative process, the committee hearings gathered more detailed estimates for the cost of IT upgrades than they had in 2014:

- $230,000 for IT upgrades at Department of Public Health
- $2 million per year for Department of Public Health to correct existing records
- Loss of revenue of $450,000 per year to Department of Public Health since they would not be able to electronically transmit names to SSA (at $3 per name) containing diacritics
- Up to $12 million for local governments to upgrade their systems
- $1–3 million in upgrades to Department of Health Care Services
- Unknown administrative costs to Department of Social Services

The sticking point for Governor Brown was compatibility with federal databases, which do not accept diacritics. In his veto message, he argued that the risks to vital records outweighed the benefits of cultural openness:

"Mandating the use of diacritical marks on certain state and local vital records without a corresponding requirement for all state and federal government records is a difficult and expensive proposition. This bill would create inconsistencies

in vital records and require significant state funds to replace or modify existing registration systems."

The committee findings make it clear that the state would incur nontrivial costs to update the name registration systems. But no mention is made of the possibility of finding creative solutions to the problem of encoding diacritics. Even assuming that government systems cannot be made to support hte full UTF-8 standard, there are ways of representing information using ASCII. For example, the international specification for machine-readable passports has a variety of control sequences for representing subtle distinctions in the Latin, Cyrillic, and Arabic alphabets using only the 26 plain characters of the English alphabet. The original reform can be recovered nearly losslessly using the transliteration table.

3.2 Massachusetts

In Massachusetts, the "characters have to be on the standard american [sic] keyboard. So dashes and apostrophes are fine, but not accent marks and the such". [Larson11]

3.3 New Hampshire

"All special characters other than an apostrophe or dash" are prohibited. [Larson11]

3.4 Kansas

Restrictions are similar to those in Massachusetts. [Larson11]

4 Passports

The standard protocol for machine-readable travel documents is Document 9309 issued by the International Civil Aviation Organization. Since all countries must operate on shared standard, it is necessary to forge a compromise between cultural openness and international security. The ICAO-9309 standard gives significant flexibility to the inclusion of national characters.

The passport data page is divided into two sections: the Visual Inspection Zone (VIZ) and the Machine Readable Zone (MRZ). Countries may fill the required fields of the VIZ however they desire in the national language, provided a transcription or translation is also provided into English, Spanish, or French. Thus

compliant passports do not per se coax a country toward the adoption of standard alphabets while still allowing international cooperation.

In the Machine Readable Zone, a transcription into the organization's approved ASCII subset (0-9, A-Z, <) is required, for the purpose of machine recognition. For Latin alphabet languages, most characters containing diacritics simply have the mark dropped, although some characters have recommended control sequences to losslessly transliterate the character. The document spells out a more extensive scheme for Cyrilic and Arabic characters which allows nearly lossless recovery of the original form from the highly schematic MRZ specification. There is even a sample Python program for converting from the MRZ to Unicode Arabic.

4.1 Latin

In the ICAO's recommendation, a distinction is made regarding the relative salience of diacritical marks; those such as the acute or grave accents over vowels are simply eliminated in the MRZ. However, they provide methods of encoding more salient characters such as the German umlaut vowels (ä,ö,ü) or the Spanish tilde ñ.

So the name "Térèsa Cañón" would become CANXXON«TERESA in the MRZ. Likewise, "Wilhelm Furtwängler" would become FURTWAENGLER«WILHELM. (b.4.2) It would not be difficult to expand the escape sequence system in order to represent additional diacritical marks.

4.2 Cyrillic

The ICAO transcription system permits a one-to-one transcription between the machine-readable representation and the original language. The system even recognizes the different values that a Cyrillic glyph might take in various languages. For example the letter Ю is transliterated as "IU", unless it is the first character of a Ukrainian name, in which case "YU" is permitted. Likewise for Щ; this is SHCH, except in Bulgarian, where it is SHT.

4.3 Arabic

For example, the Arabic name

ابو□ر□ب محمد ابن□ر□ز□الراز

would be rendered in the MRZ as:

ABW<BKR<MXHMD«BN<ZKRYA<ALRAZY

While this looks almost incomprehensible when read by a human, the use of X as an escape character allows a one-to-one transliteration back into the original script. More examples below:

B.5.9 Further examples

Arabic: هاري الشماع
VIZ: **Hari Al-Schamma**
MRZ: HARY<ALXSHMAE<<<<<<<<<<<<<<<<<<<<<<<<<<

Arabic: سمير بادمكدوذيل
VIZ: **Samir Badmakduthal**
MRZ: SMYR<BADMKDWXDHYL<<<<<<<<<<<<<<<<<<<<<<

Arabic: جمال عبد الناصر
VIZ: **Gamal Abdel Nasser**
MRZ: JMAL<EBD<ALNAXSSR<<<<<<<<<<<<<<<<<<<<<<<

Arabic: العباس عبد الله بن محمد السفاح
VIZ: **al-'Abbās 'Abdu'llāh ibn Muhammad as-Saffāh**
MRZ: ALEBAS<EBD<ALLXH<BN<MXHMD<ALSFAXH<<<<<<

Arabic: عبدالله محمد بن عمر بن الحسين فخر الدين الرازي
VIZ: **Abdullah Muhammad ibn Umar ibn al-Husayn Fakhr al-Din al-Razi**
MRZ[7]: EBD<ALLXH<MXHMD<BN<EMR<BN<ALXHSYN<FXKHR

Arabic: عبدالعزيز بن متعب
VIZ: **Abdul Aziz bin Mithab**
MRZ: EBD<ALEZYZ<BN<MTEB<<<<<<<<<<<<<<<<<<<<

Arabic: إسماعيل عزّ الدين
VIZ: **Isma'il Izz-ud-din**
MRZ: ISMAEYL<EZZ<ALDYN<<<<<<<<<<<<<<<<<<<<<

Arabic: جميلة نعيمة
VIZ: **Jamillah Na'ima**
MRZ: JMYLXAH<NEYMXAH<<<<<<<<<<<<<<<<<<<<<<