Comparing Relative Typing Difficulties Across Languages Using Corpora

Hugh J. Paterson III

Collaborative Researcher

i@hp3.me

We present the results of an experiment which compares the relative effort required to type (text input) a communicative message by using parallel texts (in different languages). Theoretically, parallel texts convey the same amount of information; therefore, they present an idealized way of comparing communicative throughput across languages. Additionally, laptop keyboards (with the exceptions of ANSI, JIS, and ISO key quantities) have the same basic physical shape, leading to reduced variability in text input options. Orthographic patterns in languages produce hotspots of action on keyboards which may be considered by typists as "difficult" — "zq" is a rare English combination, but if typed would constitute a difficult fingering on an English keyboard layout. Several anecdotal remarks exist in the language documentation literature (Boerger 2007, [ntu]; Cooper 2005:160, [kls]; Guérin 2008, [mkv]; Jany 2010, [pxm]) suggesting that "hard-to-type" motivates some ethnolinguistic-minority communities to seek orthography reform. While this perception may be community-based, a critical question is overlooked when orthography reform is the presumed solution. That is, to what extent is the keyboard layout a problem in orthography usage and therefore responsible for impressions like "hard-to-use"? As far as we know, no method has yet been proposed in which the work of typing the same text across different languages could be compared. That is, would a certain text be easier to type in English [eng], French [fra], or Eastern Dan [dnj]? This is especially relevant for language development work in multilingual contexts where technology users have the choice between using two languages in a digital context. To address this question we turn to parallel corpora and keyboard optimization algorithms. However, instead of optimizing keyboards, we use the algorithms to provide keyboard-corpus pairings a score and then compare the scores. We use a parallel corpus of English, French, and Eastern Dan (Ivory Coast). Two are non-tonal; the third tonal. The parallel corpus includes an orthography with no diacritics (English), one with some diacritics (French), and one with many diacritics (Eastern Dan). Our results suggest that current text-input methods which

rely on deadkey combinations significantly increase the text input difficulty for languages which have usage-based needs outside of the ASCII range. This suggests reconsidering previously reported research (Feit, et al. 2016) no major differences were found in typing patterns employed by users of different languages (English and Finnish).

Abbreviations

ANSI = American National Standards Institute
ASCII = American Standard Code for Information Interchange
ISO = International Standards Organization
JIS = Japanese Industrial Standards (Committee)

References

Boerger, Brenda H. 2007. Natqgu Literacy: Capturing Three Domains for Written Language Use. *Language Documentation & Conservation* 1(2). 126–153. http://hdl.handle.net/10125/1601

Cooper, Gregory. 2005. *Issues in the Development of a Writing System for the Kalasha Language.* Sydney, Australia: Macquarie University dissertation.

Feit, Anna Maria, Daryl Weir & Antti Oulasvirta. 2016. How We Type: Movement Strategies and Performance in Everyday Typing. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* 4262-73. San Jose California USA: ACM. doi:10.1145/2858036.2858233

Guérin, Valérie. 2008. Writing an endangered language. *Language Documentation & Conservation* 2(1). 47–67. http://hdl.handle.net/10125/1804

Jany, Carmen. 2010. Orthography Design for Chuxnabán Mixe. *Language Documentation & Conservation* 4(1). 231-53. https://scholarspace.manoa.hawaii.edu/handle/10125/4481