Summary: The digital technology industry is facing a trade-off: the five-percent gamble. By supporting five percent of all spoken languages in its popular applications and tools it reaches most of the world's population, yet ninety-five percent of all languages are left out. This dilemma requires innovative solutions. When decisions made by the digital technology industry carry more weight for language development than those of nation-states, then providing proper language support must be regarded as more than a business or market decision, and as a fundamental aspect of media-related governance.

Kurz gefasst: Die Digitalindustrie steht vor einem Zielkonflikt: dem Fünf-Prozent-Wagnis. Indem sie fünf Prozent aller gesprochenen Sprachen in ihren populären Anwendungen unterstützt, erreicht sie zwar das Gros der Weltbevölkerung. Jedoch bleiben dabei 95 Prozent aller Sprachen außen vor. Für diesen Zielkonflikt sind innovative Lösungen gefordert. Denn wenn der Fortbestand gesprochener Sprachen noch stärker von der Digitalindustrie beeinflusst werden kann als von nationalen Staaten, dann darf die digitale Unterstützung von Sprachen nicht nur Business- und Marktentscheidungen unterliegen, sondern muss zu einem wichtigen Aspekt von medienbezogener Governance werden.

Governing a polyglot Internet How decisions taken by the digital technology industry shape the future of languages

Thomas Petzold and Han-Teng Liao

We are witnessing a historical infrastructure shift: from a monolingual-only to a multilingual-ready Internet. Take the Web giants Wikipedia and Google as examples: They currently support around 300 languages. Compare this number to traditional broadcasting media, such as the BBC World Service currently serving 28 languages. Nonetheless, 300 is only around five percent of the over 6,500 languages in the world. Clearly each language and its speakers benefit unequally from this multilingual shift. We call this situation a five-percent gamble made by the digital technology industry on global information markets. The gamble is to assume that it is sufficient to reach the majority of the world's population by supporting five percent of the world's languages. Or, to put it more bluntly, to cover

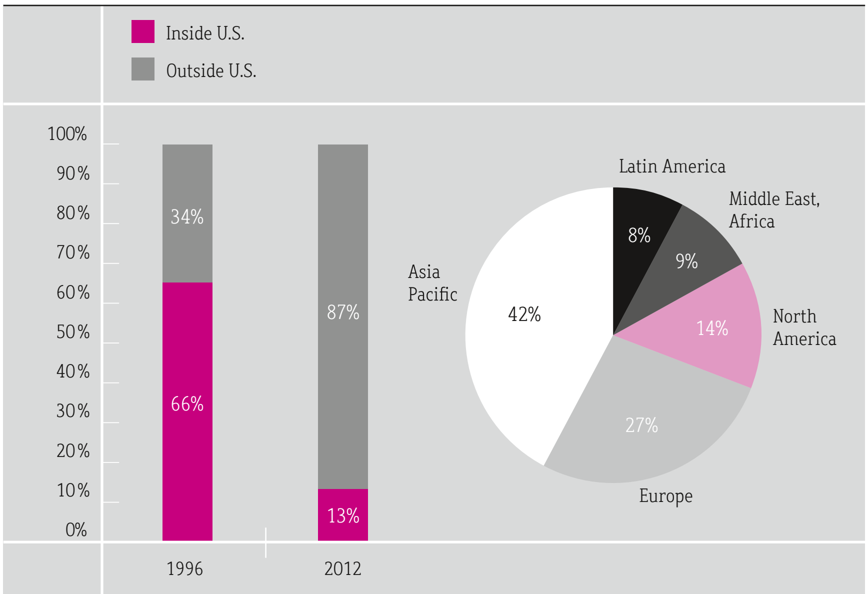


Figure 1
The evolution of the Internet population, and its current distribution (December 2012; comscore 2013)

most of the existing markets and users worldwide. Thus more than ninety-five percent of the world's languages remain unsupported.

To improve the situation, we need further social and technical innovations to enable better provisions and governance of language support in the digital environment.

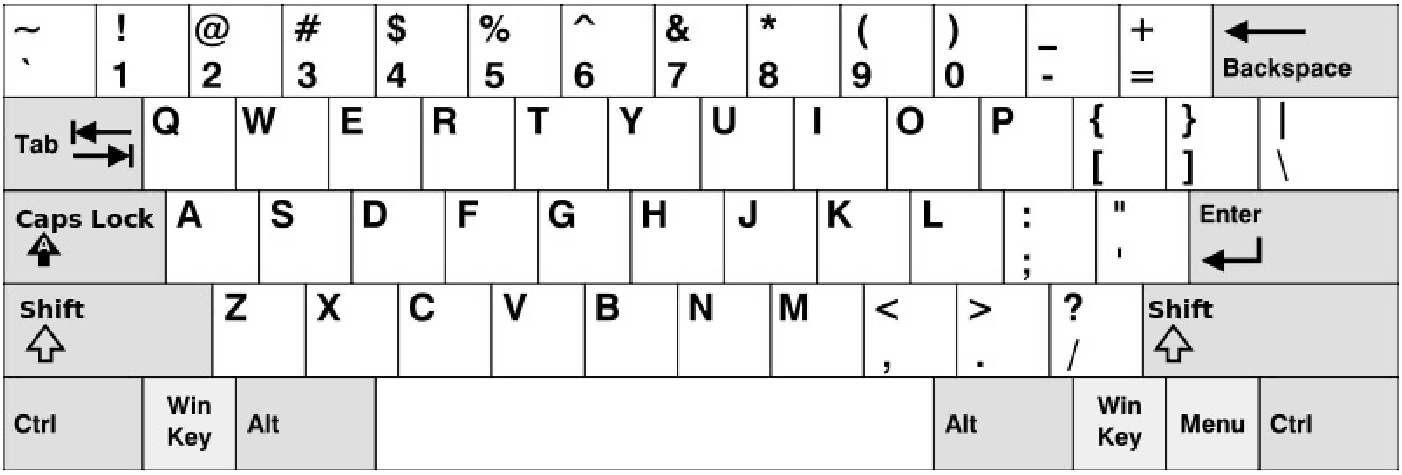
To be fair to the industry, the five-percent gamble has allowed more people outside of the US to become new Internet users, the majority of whom demand support for languages other than English. Figure 1 shows not only the contrast in the world's Internet population between 1996 and 2012, but also the most recent reality: almost 90 percent of all Internet users live outside of the US, with a large proportion of Internet traffic in 2012 generated outside North America.

The five-percent gamble marks a historical achievement in internationalizing the Internet, an American invention. Before the mono-to-multilingual infrastructure shift occurred, an American linguist asked the question: "Will the Internet always speak English?" Many politicians and entrepreneurs followed by criticizing the prevalence of English in the early Internet environment. The French president at the time, Jacques Chirac, spoke of a "major risk for humanity" which would eventually lead to linguistic and cultural uniformity. The director of a Russian Internet provider pointed to the existing lack of Russian language content and described the resulting difference between English-speaking and non-English-speaking users as "the ultimate act of intellectual colonialism." English has been the language-of-power from the beginning, but quite a contested one.

Recognizing the issue in the late 1990s, the Internet Engineering Task Force (IETF)—the main international technical community dealing with Internet architecture and related standards—declared that the "Internet is international" and thus it is "an absolute requirement to interchange data in a multiplicity of languages which (...) utilize a bewildering number of characters." Standardization efforts in encoding (that is, representing language symbols in computer digit codes) were the key driver for such a development. Encoding standards are crucial for the digitization of language texts. The main coordination body was a non-profit organization called the Unicode Consortium, with full members coming primarily from major computing and Internet companies. The consortium's work resulted in an international industry standard called the Unicode Standard, to deliberate and deliver a universal character set for every language in the world. The Unicode Consortium also maintains a unique repository that provides key components for building software that helps to automatically define the user's language, country, local date and time, local currency and other settings. These specifications are relevant for major software and Internet companies so that they can provide software and Web services which meet the different language, regional and technical requirements of a target market.

Thomas Petzold is a social technology analyst, TED speaker and professor of media management at HMKW – University of Applied Sciences for Media, Communication and Management in Berlin, Germany. As a research fellow at the WZB (2011–2013), he led a project on languages and big data in social technology. *[photo: David Ausserhofer]*

t.petzold@hmwk.de

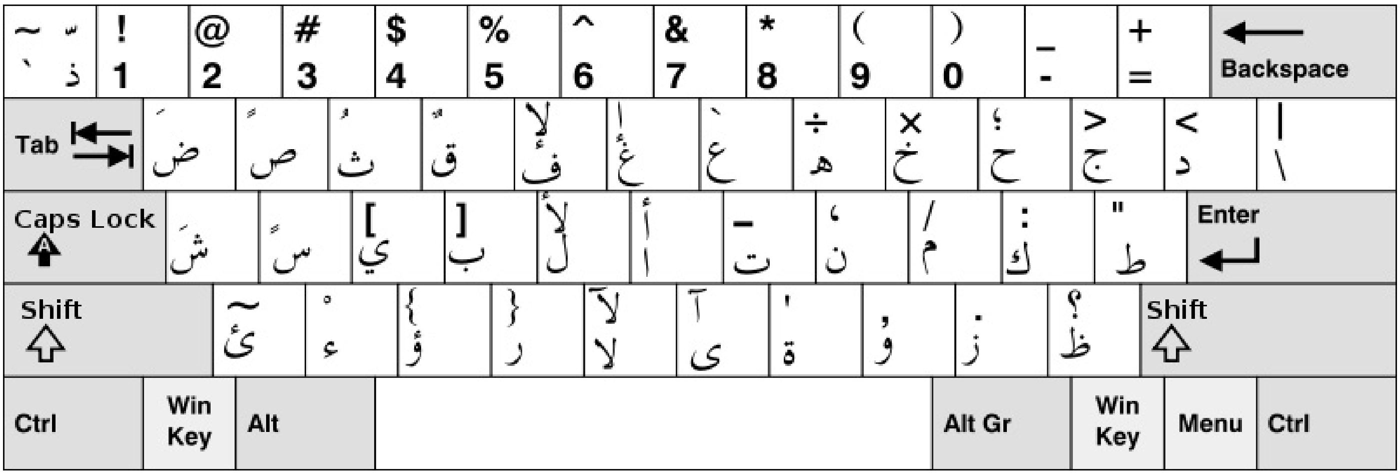

~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 _ - + = Backspace
Tab Q W E R T Y U I O P { [}] | \
Caps Lock A S D F G H J K L : ; " ' Enter
Shift Z X C V B N M < , > . ? / Shift
Ctrl Win Key Alt Alt Win Key Menu Ctrl

Besides the aspect of encoding, there are other requirements to allow users to read and write in the digital environment. Take text processing as a tangible example, where learning to use keyboards is essential. The figures show the respective layouts of an American English, Arabic and Taiwanese Chinese keyboard. Since English had a head start regarding character allocation on keyboards (which dates back to the age of the typewriter), other languages have to adapt to and repurpose an existing industry standard (also known as QWERTY). These other languages basically need to cram their characters on keyboards to be digitally ready.

Because of the investment needed in language support and digital literacy, the five-percent gamble can be seen as the direct outcome of the Return on Investment calculated by the software and Internet industry in the overall context of internationalization and localization. Internationalization of a certain piece of software means that its linguistic and regional aspects are designed and developed without regional limitations: parameters that define language and geographic aspects. Thus, internationalization prepares a piece of software to be independent from configurations of a certain language and/or region, so that it can be repurposed to serve other languages and/or regions. Localization, on the other hand, refers to repurposing already-internationalized technologies for specific locations. The internationalization process ensures that a piece of software is built language-neutral. The localization process then allows implementation of different kinds of language- and region support. Viewing these achievements, the five-percent gamble marks an important step towards multilingual support.

However, the benefits delivered and received by different language users vary greatly. The cost-benefit analysis of language support and digital literacy favors either languages that are relatively cheaper to support, for example languages using Latin alphabets, or languages that have huge market benefits, such as Chinese and Arabic. Thus, the absolute requirement to interchange data in a multiplicity of languages is at best only partially fulfilled by the industry. A considerable gap remains. It is, perhaps, also for this reason that Google recently replaced the term "knowledge" with "information" in its mission statement: "... to organize the world's information and make it universally accessible and useful." The current trade-off between 'knowledge diversity' and 'market efficiency' is made at the expense of the former and in favor of the latter.

As decisions made by the digital technology industry at times carry more weight on language development than nation-states, providing proper language support should not be regarded as simply a business or market decision, but rather as an important and fundamental aspect of media/Internet governance. For instance, aiming to digitally represent the Universal Declaration of Human Rights (UDHR) in multiple languages, the UDHR in Unicode project emphasizes the dimension of linguistic and cultural rights of language support online. Or, handling user proposals to open or close a language project, the Wikimedia Foundation Language Committee demonstrates an alternative language governance



model beyond cost-benefit market strategies: a do-it-yourself model that requires active user contribution. These practices provide valuable grounds for social and technical innovation in alternative decision-making processes of digital language support.

The digital technology industry should not be satisfied with the current five-percent gamble. At the very least, they should allow more users of languages of small populations to serve themselves with user-contributed content, tools and social networking for the sake of technological capacity-building for a better "international platform" than that of competition. In addition, online language support and digital literacy efforts can be reframed as invaluable public service or corporate social responsibility. Cultural institutions such as galleries, libraries, archives and museums may invest more in better language and content support to bridge the gap between institutional content and user-generated content online. Overall, opportunities are abundant for both private and public players to try innovative social and technical measures, for instance through crowd-sourcing and open standards/technologies to serve more users in more meaningful ways. Thus language barriers are turned into valuable resources. We must recognize that the provisional state of the current multilingual Internet environment is neither satisfactory nor innovative enough to unleash the vast potential of human knowledge. We need to utilize the full potential of an unprecedented infrastructure for language support.

Han-Teng Liao is a summer fellow at the Alexander von Humboldt Institute for Internet and Society (HIIG) and a doctoral candidate at the Oxford Internet Institute (OII). His research focus is on user-generated content and data, Web analytics (webometrics), Chinese Internet Research and integrated digital research designs (both qualitative and quantitative).
[photo: private]

hanteng@gmail.com

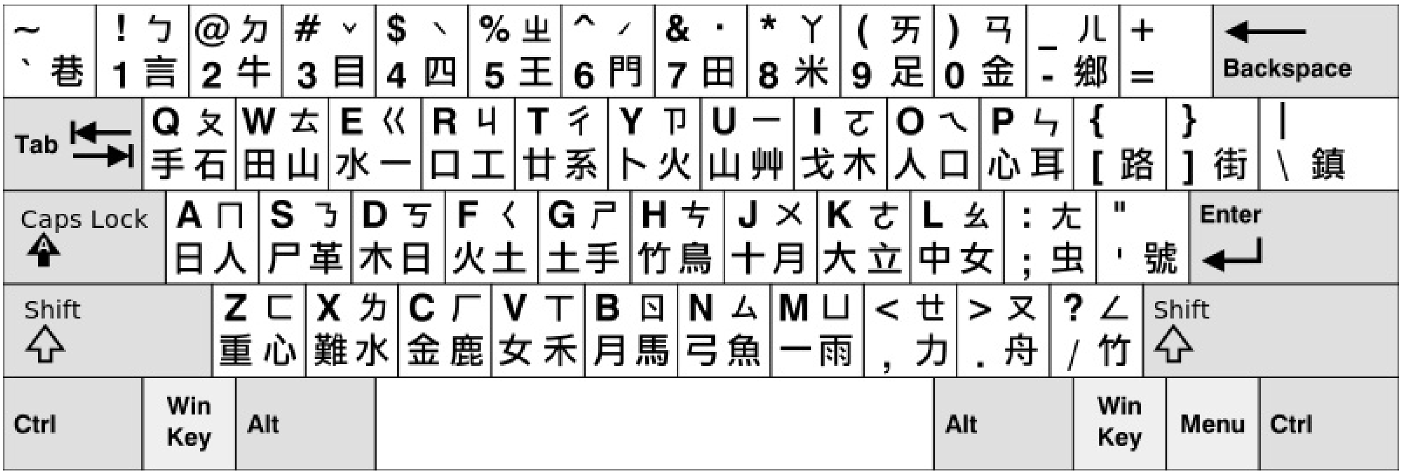
Literature
Nunberg, Geoffrey: "Will the Internet Always Speak English?". In: The American Prospect, 2002, November 30, online: http://bit.ly/1c2ujhX (accessed August 2013).

Petzold, Thomas: "36 Million Language Pairs – How to Unleash the True Momentum of Knowledge". TEDx Berlin, 23 November 2012, online: http://bit.ly/17tsW4b (accessed August 2013).

Liao, Han-Teng: Needing to Have a Voice: Linguistic Grouping in the Digital Networked Environment. ISD Working Papers in New Diplomacy, Institute for the Study of Diplomacy. Washington, DC: Georgetown University 2011, online: http://isd.georgetown.edu/files/Needing%20to%20Have%20a%20Voice.pdf (accessed August 2013).

Specter, Michael: "Computer Speak; World, Wide, Web: 3 English Words". In: The New York Times, 1996, April 14, online: http://nyti.ms/cA2mt (accessed August 2013).
The Unicode Consortium: About the Unicode Standard, Version 6.2.0. Mountain View, CA: The Unicode Consortium 2012, online: http://bit.ly/6S7mLO (accessed August 2013).

Dunne, Keiran: Perspectives on Localization. American Translators Association Scholarly Monograph Series XIII. Amsterdam & Philadelphia: John Benjamins Publishing 2006.


~ ` 巷
! 1 ㄅ 言
@ 2 ㄉ 牛
# 3 ˇ 目
$ 4 ˋ 四
% 5 ㄓ 王
^ 6 ˊ 門
& 7 ˙ 田
* 8 ㄚ 米
(9 ㄞ 足
) 0 ㄢ 金
_ - ㄦ 鄉
+ =
Backspace
Tab
Q ㄆ 手石
W ㄊ 田山
E ㄍ 水一
R ㄐ 口工
T ㄔ 廿系
Y ㄗ 卜火
U ㄧ 山艸
I ㄛ 戈木
O ㄟ 人口
P ㄣ 心耳
{ [路
}] 街
| \ 鎮
Caps Lock
A ㄇ 日人
S ㄋ 尸革
D ㄎ 木日
F ㄑ 火土
G ㄕ 土手
H ㄘ 竹鳥
J ㄨ 十月
K ㄜ 大立
L ㄠ 中女
: ㄤ ; 虫
" ' 號
Enter
Shift
Z ㄈ 重心
X ㄌ 難水
C ㄏ 金鹿
V ㄒ 女禾
B ㄖ 月馬
N ㄙ 弓魚
M ㄩ 一雨
< ㄝ , 力
> ㄡ . 舟
? ㄥ / 竹
Shift
Ctrl
Win Key
Alt
Alt
Win Key
Menu
Ctrl