

36 Million Language Pairs: Generative Multilingualism in Digitally-Enabled Societies

Thomas Petzold¹

ARC Centre of Excellence for Creative Industries and Innovation,
Creative Industries Faculty,
Queensland University of Technology, Australia
t.petzold@qut.edu.au

Abstract This essay is based on a study that explores the relationships among multilingualism, technological change and the generation of novelty. It discusses the development of a dominant paradigm in language governance within digital culture and goes on to introduce the notion of ‘generative multilingualism.’ Generative multilingualism, it is argued, regards it as insufficient to foster the inclusion of most or all languages on the Internet if no adequate regime of intersection amongst those languages is put in place. In order to be sustainable in digital culture, linguistic pluralism not only entails the participation of the world’s 6.000 languages, it requires their mutual interaction. The enormous task linguistic pluralism creates, therefore, is the generatively-enabled interactivity of 36 million language pairs.

Linguistic pluralism was not a central part of the Internet’s foundational structure when it was introduced forty years ago. Instead, the idea of a distributed architecture of computers was invented for the US military to have a more robust infrastructure in case of attacks or emergencies. At that time, the Internet spoke one language – English. This has fundamentally changed over the last decades with the mobile net making use of more and more languages.

Some argue that today’s Internet architecture has a significant effect on language and linguistic pluralism. Some linguists describe the impact of the Internet as ‘most revolutionary’ because, as they argue, ‘it is extremely rare to find changes which are so global that they affect all languages’ (Crystal 2004). Some studies have investigated those changes looking at language use in different contexts (e.g. Danet & Herring 2007; Crystal 2001, 2004, 2006). The public acquisition of the Internet as an additional medium of communication, they argue, affects languages in two ways: ‘it initiates change in the formal character of the languages which use it; and it offers new opportunities for languages to use it.’ (Crystal 2004: 81) The inversion of the latter constitutes linguistic pluralism in digital culture, and will be at the core of this essay, which addresses some of the pertinent questions raised by the 12th Berlin Roundtables on Transnationality background paper, in particular pertaining to linguistic pluralism, language regimes and the Internet.

¹ The author would like to thank the following organisations for financial support provided towards this research: German Academic Exchange Service (Bonn, Germany), Kurt-Tucholsky-Foundation (Hamburg, Germany), Creative Industries Faculty at Queensland University of Technology (Brisbane, Australia).

How we perceive and make use of linguistic pluralism in digital culture, I argue, depends crucially on the generative character of dominant paradigms of language governance. By that I mean it is insufficient to foster the inclusion of most or all languages on the Internet if no adequate regime of intersection amongst those languages is put in place. Taken seriously, linguistic pluralism in digital culture entails not only the participation of the world's 6.000 languages, but also their mutual interaction. Based on the Internet's sustainability as a 'generative system' ([Zittrain 2008: 70](#)), the enormous task linguistic pluralism creates is the generatively-enabled interactivity of 36 million language pairs, i.e. the inter-translatability amongst 6.000 languages ($6.001 \times 6.000 \sim 36$ million).

Therefore, confining linguistic pluralism to issues of access and participation as in debates about 'digital divide' and 'digital literacy', for example, stops short of understanding what linguistic pluralism is *for* in digital culture.² A participatory understanding can most notably explain how access to and participation on the Web is important for a specific language and its speakers (and how it is not), and why languages with fewer speakers are potentially and actually disadvantaged against major languages (and why they are not). In order to grasp the extended uses of linguistic pluralism in digital culture, however, we need to develop an understanding which builds upon and goes beyond participatory characteristics. We need to introduce the notion of linguistic generativity or what I call generative multilingualism.

Jonathan Zittrain ([2008: 70](#)) uses the term 'generativity' to describe the Internet (and some of its components) as a generative system which 'might grow or change over time as the uses of a technology by one group are shared with other individuals, thereby extending the generative platform.' Generativity expresses the ability of a self-organising system to create and generate independently, without any or only little input of the system's originator. The Internet acquired generative traits only after 'centrally orchestrated improvements by proprietary networks plateaued.' (id.: 82) This allowed for a more open development in technology which in turn 'led to developments in content and ultimately in social and economic interaction: the

² The Internet has been described as an eclectic medium, where participation by borrowing, combining and creating content is seen both as an expression of everyday creativity and as a matter of concern (see e.g. [Jenkins et al 2006](#)). Major debates about linguistic pluralism on the Internet revolve around the issue of participation, and are most notably concerned with issues such as 'digital divide' and 'digital literacy.' The former focuses on technological aspects of access to the Internet ranging from the availability of a physical Internet connection to the development of inclusive computing industry standards such as Unicode. 'Digital literacy,' on the other hand, is concerned with the way of acquiring and applying necessary skills to navigate and make use of the different modes of expression in digital culture ([cf. Hartley 2009](#)). Linguistic pluralism is affected by both instances of participation.

Web and Web sites, online shopping, peer-to-peer networking, wikis, and blogs.’ (loc. cit.) The strength of a generative system, Zittrain argues, is its ability to trigger unanticipated change (‘innovative output’) and to allow for the inclusion of large and varied audiences (‘participatory input’). Generative tools are usually developed across a wide range of individuals and enterprises. They are thus, on the one hand, individually useful and, on the other, ‘more basic’ and ‘less specialized’ for the accomplishment of specific services. This has had repercussions for those who participate in those environments. Moreover, these developments are significant for linguistic pluralism in generatively-enabled societies and economies.

One example in this context is the use of Wikipedia (as one instance of a generative tool) in languages with fewer speakers, such as the Sorbian Wikipedia. Sorbian is a tiny remnant of a multitude of Slavic dialects and languages widely used from the sixth century onwards. Today, it is spoken in the Lusatia region in far-Eastern Germany, south of Berlin, by around 60.000 people.³

Sorbian is part of the remit of public-service broadcaster MDR (Mitteldeutscher Rundfunk), which has provisions for radio and television programs in both Upper Sorbian (spoken mainly in the districts around Bautzen, or Budyšin) and Lower Sorbian (spoken mainly in the districts around Cottbus, or Choćebuz). MDR’s Sorbian broadcasts cover several hours a day but it is not an around-the-clock service. The Sorbian Wikipedia (Upper and Lower Sorbian versions), as an additional means of communication running in a 24/7-modus operandi, thus provides a novel forum for reading entries, engaging in discussions, and enjoying audiovisual content. This online forum allows, at least in principle, for all three of the distinctive functions commonly associated with public service broadcasting: to inform, educate and entertain.

The more successful Upper Sorbian Wikipedia consists of around 5.800 articles, more than 137.000 edits, around 3.800 users (of which around 50 are considered as active) and three administrators (as of 01/2010). If measured by article counts the Sorbian version ranks 107 out of more than 270 different language versions. Article count is, however, only one statistical element for measuring changes of Wikipedia versions, and a controversial and insufficient one as well (cf. [Jones 2009](#)). An alternative statistical indicator would balance the power and possibilities of a language, measured by editors per million language speakers, for

³ Different sources estimate the number of Sorbian speakers who use the language as an everyday medium of communication between 12.000 and 30.000.

example. If taken as a stand-alone measurement it sees the Sorbian language versions being ranked amongst the Top-20 Wikipedia versions (cf. Figure 1).

Wikipedia Statistics								
Thursday December 31, 2009								
Wikipedias		See bottom of page for language comparisons / other reports			W = Wikipedia Article		= Charts	
Languages			Regions	Participation			Usage	Content
Article	Code	Name		Speakers in millions (log scale) (?)	Prim.+Sec. Speakers	Editors (5+)	Views	Article
Charts	Project	Tables		Editors per million speakers (5+ edits)	M=millions k=thousands	per million speakers	per hour	count
Σ All languages (270)			AF AS EU NA SA OC CL W					
W	vo	Volapük	CL		10	700000	4,210	118,785
W	kw	Cornish	EU		245	16327	432	1,813
W	pih	Norfolk	OC		600	3333	93	271
W	stq	Saterland Frisian	EU		2 k	2500	206	1,478
W	gv	Manx	EU		2 k	2353	655	2,953
W	an	Aragonese	EU		10 k	1600	1,790	18,571
W	mwl	Mirandese	EU		15 k	400	222	848
W	dsb	Lower Sorbian	EU		14 k	286	352	717
W	se	Northern Sami	EU		20 k	200	384	2,825
W	is	Icelandic	EU		320 k	150	3,157	27,073
W	fo	Faroese	EU		70 k	143	664	3,870
W	no	Norwegian	EU		5 M	136	39,365	238,657
W	szl	Silesian	EU		56 k	125	341	1,915
W	hsb	Upper Sorbian	EU		40 k	125	581	5,811
W	fi	Finnish	EU		6 M	118	71,049	226,029
W	rm	Romansh	EU		35 k	114	240	3,185
W	eo	Esperanto	CL		1 M	105	14,375	123,301
W	sv	Swedish	EU		10 M	103	88,399	341,443
W	br	Breton	EU		250 k	100	2,716	31,157
W	sa	Sanskrit	AS		50 k	100	284	3,910

Figure 1: Top-20 Wikipedia versions by editors per million speakers; Source: stats.wikimedia.org

Such a measurement obviously biases languages with fewer speakers. If not put into context or used together with additional indicators, such singular statistical snapshots do not provide for a sustainable future when merely used to diverge from serious challenges of those languages.⁴ A language with fewer speakers, it is assumed, survives as long as sufficient speakers guarantee essential vitality and interaction. The subsequent causal deduction has been, at least in principle, that once such viable interactivity stalls the main resource for the identity and integrity of all individuals concerned ceases to exist, with the global heritage accordingly affected. This is no longer a valid general assumption if linguistic pluralism and languages with fewer speakers accomplish comprehensive integration into a generative system. This is one important aspect and consequence of linguistic generativity.

Linguistic generativity on the Internet is dependent on a variety of parameters. Against this background, it is important that both the architecture of the Internet as well as some of its

⁴ The most significant issues amongst those for languages with fewer speakers are loss of editors and quality of articles (cf. van Dijk, 2009).

major tools feature visible traits of a generative system. Furthermore, as Arthur (2009: 209) argues, the way the economy works is also increasingly generative, by ‘shifting from optimizing fixed operations into creating new combinations, new configurable offerings.’ This is significant for the development of linguistic pluralism on the Internet if we consider some of the key players and their respective paradigms of language governance in digital culture. At the moment, we can distinguish at least four dominant models of language governance, exemplified here by ICANN, Wikimedia, SBS Australia and Google. Figure 2 illustrates those key players and their respective logic of language governance pertaining to three major parameters: community, politics, and market.

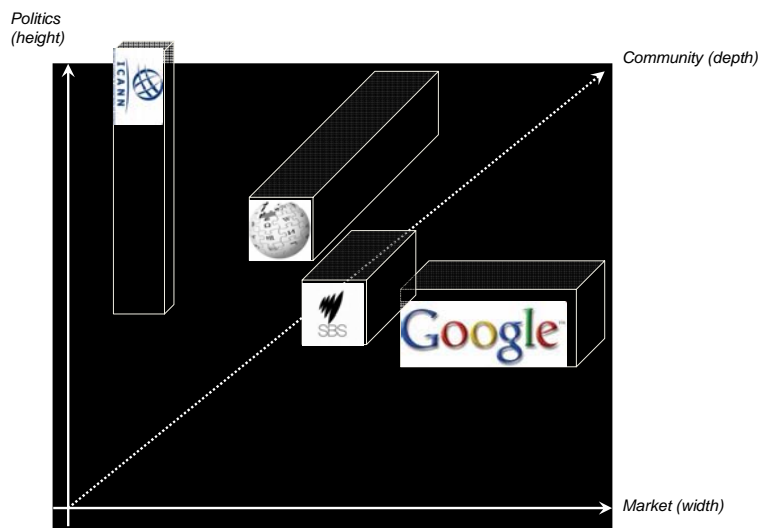


Figure 2: Dominant Agents of Language Governance in Digital Culture

The following section provides a brief overview of these key agents discussing how they affect linguistic pluralism in digital culture:

1) **ICANN**, the Internet Corporation for Assigned Names and Numbers, is the first and largest Internet governing body to date and was created under the auspices of the US Department of Commerce as a non-profit organisation in 1998. Much of its work on language governance has been concerned with the introduction of so called Top-Level Domains (TLDs, e.g. the .de in <http://www.example.de>). As is emphasized in figure 2 the ICANN model of language governance is highly political, and may best be described as *gift logic*. For example, a recent decision to introduce non-Latin characters⁵ in Top-Level-Domains was described as the biggest change to the way the Internet functions since its

⁵ Languages containing non-Latin characters include but are not limited to Arabic, Chinese languages, Cyrillic, Central and Eastern European languages, Hebrew, Greek and Japanese.

inception 40 years ago ([BBC 2009](#)). This is generally correct. The decision, however, was delayed for half a decade and, according to Paul Hoffman, who is one of the authors behind the original development of the technology for non-Latin characters, it still has its limitations:

ICANN's announcement 'only' covers IDNs⁶ in country names, not in new TLDs. The latter will (or won't) happen a few years from now, and the political discussions there will be even more difficult than it was for the country names. ([Hoffmann 2009](#))

In the wake of signing a new Memorandum with the US Department of Commerce ICANN has pledged more accountability, public disclosure and participation. However, low transparency and decision-making behind closed doors by paying members (annual fee: US\$185.000) still affect the way linguistic pluralism develops in the domains of ICANN's responsibility.

2) **Wikimedia**'s language governance model is in many respects the antithesis to the ICANN model and may best be described as *community logic*. Open source software and self-governance are at the core of Wikimedia's operational structure. This is reflected in the Wikimedia Foundation's language governance policy. A multiple-step process determines the addition or rejection of another language version to Wikipedia (the most commonly known of the Wikimedia projects). The application procedure is entirely documented and can be monitored at all times.⁷ The specific requirements that any new language proposal must meet before it can be approved are:

- 1) A new language edition must not already exist on any Wikimedia project.
- 2) The language must have a valid ISO-639 1-3 code.⁸
- 3) The language must be sufficiently unique that it could not coexist on a more general wiki.⁹

⁶ Internationalized Domain Names (IDN) is an Internet standard which contains at least one language-specific script or alphabet. One example for that is the use of an umlaut as in www.hörby.se.

⁷ Cf. Wikimedia's language proposal policy at <http://tiny.cc/B93yC>

⁸ This means it must be listed in an ISO-639 database, or standards organizations must be convinced to create an ISO-639 code for a 'new' language.

⁹ This, in most cases, excludes regional dialects and different written forms of the same language. It must be noted that determining the requirements of uniqueness of a new language also implies risks of a political overspill into Wikipedia. Although the Wikimedia Foundation clarifies that the uniqueness of a language will be decided individually ('considered on a case-by-case basis') and in a neutral manner ('does not consider political differences'), Wikimedia's language governance may still be indirectly affected by political disputes around definitions of 'languages' and 'dialects' that pertains ISO decision-making processes, for example.

- 4) A sufficient number of living native speakers form a viable community and audience.¹⁰

The Wikimedia model of language governance uses a traceable and, indeed, generative adoption procedure heavily involving the community. This has allowed Wikipedia to become the most linguistically diverse project in digital culture. Yet, all available Wikipedia languages taken together only account for around five percent of the world's languages. Adding new languages and improving the quality of existing language editions is one aspect of a comprehensive generative multilingualism which, in addition, also requires the interactivity amongst languages.¹¹ Generative multilingualism has the propensity to focus on sustainable modes of such interactivity which, in turn, allow for knowledge to become richer in substance, more contestable, and more useful for its large and varied audiences. Understanding and organising knowledge and information in such a manner can be significant for areas spanning from global cultural heritage to regional economies.

It is important to note that there is no single model or institution which is capable of accomplishing generative multilingualism on its own. Generative multilingualism is an entirely collaborative effort. This is the fundamental difference from earlier setups of notable institutions which foster and support multi-language use. One such example is Australia's public service broadcaster SBS.

- 3) The logic of language governance applied by **SBS Australia** may best be described as *curator logic*. The Special Broadcasting Service (SBS) is a largely publicly funded broadcaster established in the late 1970s to better reflect Australia's multicultural and multilingual society. It is a prime example of mediated multilingualism in an environment where high ambitions and scarce resources have to be constantly negotiated. In such

¹⁰ This requirement must be proven in order for the language request to receive final approval. Therefore, an active test project on a specified Wikimedia space must have been initiated where interest by individual speakers or supporters of the language is registered and arguments for and against the admission of the new language is gathered. A decision in accordance with the language proposal policy will then be made by the language committee whose goals, responsibilities, members and decisions are openly accessible at http://meta.wikimedia.org/wiki/Language_committee. In order to receive full approval the interface of the new project must be translated into the new language, such as 'Wikipedija' for Wikipedia in the Sorbian edition.

¹¹ One example for that is the analysis and organisation of *inter-language linking*. Inter-language linking describes the intersection of knowledge and goods amongst different languages. This includes, for example, the different practices of preparing, displaying, using and exchanging information like Wikipedia entries. Such interactivity can culminate, for example, in knowledge sharing or knowledge warfare. For a comparative micro-analysis of specific Wikipedia entries across languages cf. [Callahan and Herring \(2009\)](#); for a geo-linguistic macro-analysis of inter-language linking of Wikipedia entries cf. [Petzold and Liao \(forthcoming\)](#)

environments multilingualism is mediated for languages that are assumed to be ‘most suitable’ within differing political contexts (Ang et al 2009; for a critique cf. Podkalicka and Petzold 2009). Hence, SBS’ language policy depends on specific political goals set by differing political agents. And with changes in government it adjusted its overarching ‘business’ philosophy time and again.¹² Accordingly, its decision-making and language adoption procedure, i.e. the languages around which media provisions are created and scheduled for SBS’ key-constituencies, is policy dependent and not entirely transparent. Moreover, when scarcity requires the adoption of a limited number of languages and in-language programs, such models of mediated multilingualism become exposed to intensive public criticism that constrain the ability to innovate and eventually imperil the very existence of the institutions in question.¹³ The SBS model of language governance operates under a largely centrally orchestrated *modus-operandi* which allows its language communities (audiences, users) to partake more in the way ICANN does than that of Wikimedia.

Some of the major shortcomings and criticisms that models like SBS’ face in digital culture may be addressed best by taking linguistic generativity into consideration. One of the key players in linguistic generativity at the moment is Google.

4) Google’s model of language governance may best be characterized as *language market logic* where both market and linguistic considerations prevail. Google Translate, the world’s most used and equally contested translation online site, has an array of more than 50 wider and lesser-used languages amongst which can be translated. The latter trait makes Google Translate the most significant tool of linguistic generativity in digital culture as it provides for the inter-translatability between any two languages. It has the ability of generating more than the already supported 51 languages and 2.550 language pairs. It offers the utopic vision of a potential linguistic pluralism in which any of the world’s 6.000 languages can be translated into any other instantly: 36 million language pairs as the ultimate extrapolation from current possibilities.

¹² In its most recent form it focuses on ‘social inclusion,’ a cultural policy keyword used by the incumbent Rudd-government and one that has largely replaced ‘cultural diversity,’ a cultural policy keyword relied upon by the previous Howard-government.

¹³ Cf. Berlin’s Radio Multikulti, for example (discussed in Podkalicka and Petzold 2009).

Whether a new language can be added to Google Translate depends on a few criteria.¹⁴ Both Google's market economics and linguistically related technical issues need to be resolved in order to be included on Google's translation service. Moreover, Google's 'language policy' invites the initiative of others to provide sufficient bilingual textual material. Nonetheless, the request for and adoption of a new language is not as transparent as in the Wikimedia model. Whilst the process of adoption becomes reasonably understandable it is somehow unclear what is exactly involved in requesting a new language. An enquiry by the author about what amount of bilingual text is required in order to have the Sorbian languages included remains unanswered.¹⁵

Google's contribution to linguistic generativity is important. Google Translate's significance increases by becoming fully integrated into various web applications such as Google's search engine or any other tool which uses Google Translate for instant translation. This turns it from a stand-alone tool into a processing layer that is embedded in the mobile web environment, ultimately without the user even being consciously aware of the act of translation and its impact on the original text (which, in turn, raises interesting new questions and concerns).

Wikimedia's language proposal policy and Google's ability to inter-translate are at the forefront of defining a digitally-enabled linguistic pluralism based on the principle of generative multilingualism. The adoption of Wikimedia's language proposal policy by Google, it seems, would be one option for further extending the ability of linguistic generativity. Another option might be to develop an open source translation project under Wikimedia's auspices building, for example, on the Wikipedias in different languages as an initial text corpus, and the Wiktionaries for word definitions. Such a project would then go head-to-head with Google Translate. Any of these options, in turn, can be momentous for all language communities and economies partaking in digitally-enabled societies. Whilst languages may stop to exist due to its last speakers dying, the interaction between languages would never cease. This is why generative multilingualism is also important; this is what linguistic generativity is also good for. We are only beginning to see this evolving.

¹⁴ cf. Google Translate, New Languages Support section: 'We're working to support other languages and will introduce them as soon as the automatic translation meets our standards. It's difficult to project how long this will take, as the problem is complex and each language presents its own unique challenges. In order to develop new systems, we need large amounts of bilingual texts.' Accessed 15 December 2009: <http://translate.google.com/support/?hl=en>

¹⁵ Other examples of voluntary, indirect lobbying for a new language, however, have proven to be successful, as was the case for the Welsh language which became included on Google Translate in 2009 (Carlson 2009). Yet, there is still the need for a better understanding of the reasons behind the adoption of new languages in Google Translate.

References

Ang, I., Len Ang, Gay Hawkins & Lamia Dabboussy (2009) *The SBS Story: Challenges of Cultural Diversity*, Sydney: UNSW Press.

Arthur, W. Brian (2009) *The Nature of Technology: What it is and how it evolves*, New York: Free Press.

BBC (2009) *Internet Addresses Set for Change*, accessed online 30 October 2009: <http://news.bbc.co.uk/2/hi/technology/8333194.stm>

Callahan, E. and Susan C. Herring (2009) *Cultural Bias in Wikipedia Content on Famous Persons*, Presentation at Internet Research 10.0 – Internet: Critical, AoIR, Milwaukee.

Carlson, B. (2009) *Welsh “snubbed” by Google translate*, accessed online 19 September 2009: <http://blog.languagetranslation.com/public/item/231295>

Crystal, D. (2001) *Language and the Internet*, Cambridge: Cambridge University Press.

Crystal, D. (2004) *The Language Revolution*, Cambridge: Polity Press.

Crystal, D. (2006) *Language and the Internet* (2nd edition), Cambridge: Cambridge University Press.

Danet, B. and Susan C. Herring (2007) *The Multilingual Internet*, Oxford: Blackwell.

Hartley, J. (2009) *The Uses of Digital Literacy*, St Lucia: University of Queensland Press.

Hoffmann, P. (2009) *Interview with Xenji Jardin*, accessed online 2 November 2009: <http://www.boingboing.net/2009/11/02/icann-haz-cheezburge.html>

Jenkins, H. (2006) *Confronting the Challenges of Participatory Culture: Media Education in the 21st Century*, Boston: MIT Press.

Jones, C. O. (2009) *Look it up in Wikipedia*, in *Planet 195*, 27-31.

Petzold, T. and Hanteng Liao (forthcoming) *Geo-linguistic analysis of the World Wide Web: The use of cartograms and network analysis to understand linguistic development in Wikipedia*, in Araya, Daniel, T. Houghton & Y. Breindl, *Nexus: New Intersections in Internet Research*, New York: Peter Lang.

Podkalicka, A. & Thomas Petzold (2009) *Medienvermittelte transnationale Öffentlichkeiten im Zeitalter praktischer Hybridisierung* (engl. *Mediated transnational publics in environments of practiced hybridisation*), in *Nordosteuropäische Geschichte in den Massenmedien. Medienentwicklung, Akteure und transnationale Öffentlichkeit. Volume XVIII*, Lüneburg, Verlag Nordost-Institut.

Van Dijk, Z. (2009) *Wikipedia and lesser-resourced languages*, in: *Language Problems and Language Planning* 33(3), 234-250.

Zittrain, J. (2008) *The Future of the Internet – And How to Stop it*, New Haven & London: Yale University Press.

Additional Online Resources

Google Translate New Languages Support, accessed 15 December 2009:
<http://translate.google.com/support/?hl=en>