Migration of Small and Endangered Languages into the Wikipedia

Armin Hoenen
Institute for Empirical Linguistics
Goethe University
Frankfurt, Germany
hoenenarmin@gmail.com

Marc D. Rahn
Institute for Empirical Linguistics
Goethe University
Frankfurt, Germany
marc.rahn@venturerebels.de

Abstract

This paper reviews the Wikipedias for the smallest languages it is available for. We compute the most frequent shared articles (by automatically extracting their translation links), categories and other features. By analysing these data and aligning it tentatively with the literature, it is aimed at an understanding what interests could connect endangered languages. Which needs towards an emerging digital infrastructure could these results point to, is what the discussion is concerned with besides the crucial question of representativity of this data for small and endangered languages.

1 Introduction – Wikipedias as Encyclopedia

For endangered languages it is often difficult to gather enough (digital) data. Especially studies comparing multiple endangered languages are rendered difficult by this. At the same time, the internet is a platform considered vital by some for the survival of languages, compare Warschauer (2000). We choose a real world example of a website used by some endangered languages, which provides data – the Wikipedia. We take the smallest 45 Wikipedias as one example of how content of smaller and endangered languages can develop. The choice was determined by the arbitrary parameter of article number (less than 1.000) where we decided to not reduce the languages only to those clearly endangered. There were two reasons for this. Firstly, there is more than one definition of language endangerment with fluctuations on the borders. Secondly, the diversity in terms of language family, place etc. would have been reduced. A larger sample is clearly more reliable. 12 of the sampled languages are found in the UNESCO Atlas of endangered languages.[1]

The Wikipedia itself is a special instance of web content, an encyclopedia providing definitions. Definitions are an achievement of literacy (Ong, 2013, p.55). Non-literate communities are often quite place-bound (compare Ong (2013, p.46/47)) and are very aware of their surroundings. For entities within them, because of their high familiarity with them, few adults in the indigenous community (IC) would need or consult a definition. On the other hand, for things which are not or only marginally relevant for the life of the IC, a definition just as the thing itself may appear equally superfluous to the IC (compare Hoenen (2016)). This may explain a reluctance to use such platforms on the internet or slow growth rates.

However, with globalization and an ever more technological way of life, the integration of new tools and technologies into the everyday life of the IC is one engine driving contemporary change. As Domeij et al. (2019, p.119) remark, unlike other tools, the internet allows the IC a high degree of self-governance when composing web-content for their purposes: "the Internet represents a democratization, which benefits the minority languages". Fishman (1991) remarks that the survival of languages depends more on will and transmission than on actual speaker numbers and Warschauer (2000) analyzes the internet to be connected with the former of these two variables. Krauwer (2003) and Arppe et al. (2016) analyze possible language technology uses of ICs. We chose to analyze the Wikipedia as another instance of internet usage for and by ICs.

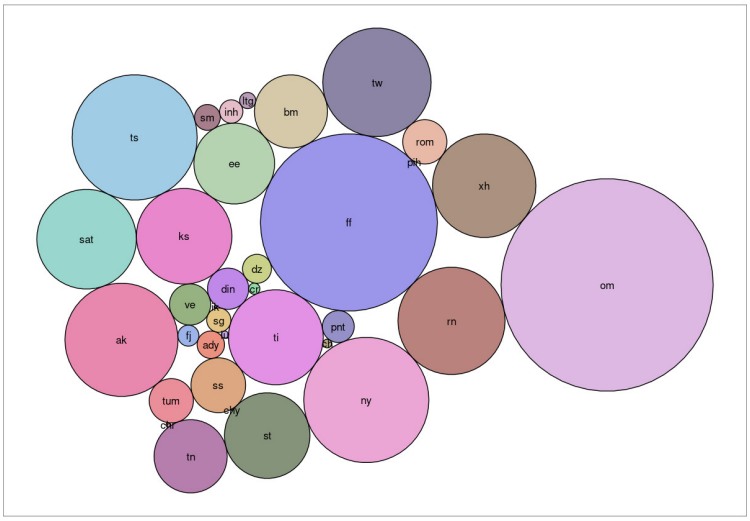
2 Languages Sampled

Languages we used featuring less than 1.000 and more than 10 Wikipedia articles (the smaller

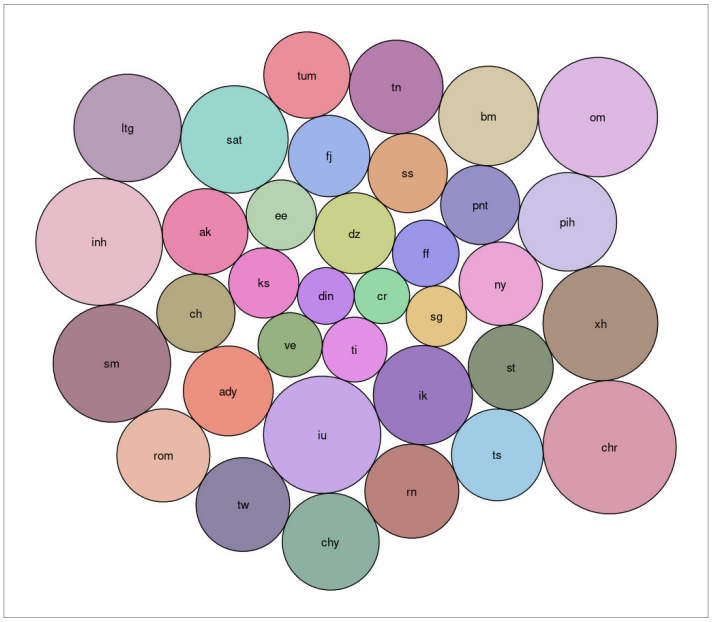
[1] We advocate a rather broad approach to endangerment which is compliant with the categorial imperative – better classify too many languages wrongly as endangered and support them than to classify too few as endangered and therefore leave some unattended.

Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages: Vol. 2 Resource Papers and Extended Abstracts, pages 41–47,
Online, March 2–3, 2021.

(1)



(2)



(3)

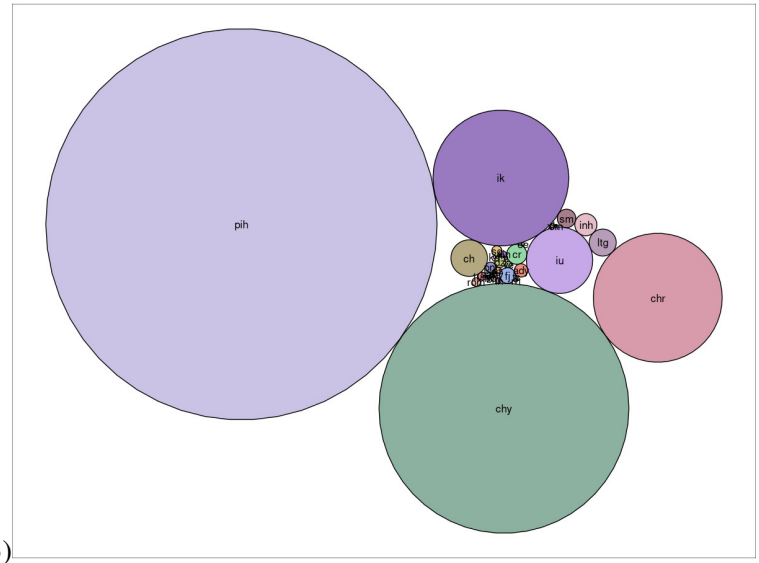


Figure 1: 1: Speaker numbers, 2: Wikipedia sizes (numbers of articles), 3: productivity (articles per speaker). Diameter with benchmark size of the largest languages as unity and the smaller ones relatively smaller - bubble diagram with R (packcircles). Iso-codes for languages.

ones are analyzed below) in March 2019 were (in alphabetical order with continent marker - A = Asia, Af = Africa, Aus - Australia & Oceania, Eu - Europe, NA - North America, SA - South America):

Adyghe (A), Akan (Af), Bambara (Af), Chamorro (A), Cheyenne (NA), Chichewa (Af), Cree (NA), Dinka (Af), Dzongkha (A), Ewe (Af), Fula (Af), Fijian (Aus), Ingush (A), Inuktitut (NA), Inupiaq (NA), Kashmiri (A), Kirundi (Af), Latgalian (Eu), Ndonga (Af), Norfolk (Aus), Oromo (Af), Pontic (A), Romani (Eu), Samoan (Aus), Santali (A), Sango (Af), Sesotho (Af), Swati (Af), Tigrinya (Af), Tsonga (Af), Tswana (Af), Tumbuka (Af), Venda (Af) and Xhosa (Af).

Largest in terms of speakers according to the Wikipedia were Oromo (34.5 million), Fula (24 million) and Chichewa (12 million); smallest Norfolk (800), Cheyenne (1900), Inupiak (6740). The data are quite diverse: our sample contains languages that use a variety of different writing systems, come from a variety of language families, from diverse locations and differ wildly in socio-economic status and speaker numbers. There is of course some 'sampling' bias, the absence of a South-American language is immediately obvious. Going deeper into this counting the number of all Wikipedia languages per continent, we find as of January 2020: Asia accounts for roughly 39%[2], Europe for 33%, Africa for 15%, North America for 5%, Australia & Oceania for 4% and South-America for only roughly 2% (which is equal to only 5 languages from 309, where one has 0 articles; another roughly 3% were artificial languages such as Esperanto, an excess of 1% in the sum stems from rounding). Now this does neither correspond very nicely to the statistics of speaker numbers per continent nor to the numbers of languages per continent as one can see from an analysis of data from ethnologue[3]. In terms of access to the internet, Eurasia and North America should be privileged while regions of high linguistic diversity such as Papua New Guinea or the Amazon basin have little access, which may skew the data additionally as well as general educational and economic variables.

Figure 1 visually displays the diversity in terms of speaker numbers and Wikipedia sizes in our sample hinting towards a per capita productivity especially elevated in some smaller languages. Norfolk (classified as vulnerable) for instance has roughly 800 speakers but the number of Wikipedia articles is roughly the same as for Kirundi which has 6 million. Being an English-based creole spoken on an island situated in the Pacific Ocean between New Zealand and Australia, access to the internet and economic well-being may help productivity. The distribution of the Wikipedia-sizes is more uniform than that of the population sizes. This may point to a general saturation, once the most important articles are being described, the growth stagnates independently of population size.

3 An Analysis of Article Titles

We download the Wikipedia Titles Dumps from the Wikimedia Foundation for all of these languages, extract them and limit all files to the entries in *namespace 0*, which is the encyclopedic main (and non meta or administration related) area. For each word, we extract the titles of the links into other languages online; these are practically translations (see also projects such as BabelNet) and process them further. Initially, we manually examined the smallest Wikipedias. In the following table, we present an overview of the articles in these Wikipedias in chronological order of their creation. The article 'main Page' has been omitted, as well as empty pages, see Table 1.

The majority of these smaller Wikis have been closed or remained in the Wikipedia incubator. The 36 larger Wikipedias contained 28,305 articles according to the titles downloaded on February 21. The mean was 786 articles per language with a minimum of 285 and a maximum of 1,646 articles.[4] Since manually assessing all of them would be out of scope of this article, we use Natural Language Processing to access some frequency information especially on categorizations of the articles.

As for the translations present (into or from Wikipedia articles on the same entity in another Wikipedia language), English accounted for most translations in all but 3 Wikipedias (Adyghe with Russian, Ingush with Russian and Latgalian with Latvian) where it came second. From those articles featuring a translation at all, an English translation was present on average in 97% across

[2]Some languages such as Russian are spoken on more than one continent, for convenience we assigned only one continent to them, usually the one with the larger assumed speaker number.

[3]https://www.ethnologue.com/guides/continents-most-indigenous-languages

[4]For the reported qualitative analyses, we excluded Gothic and Old Church Slavonic, as there is no IC behind them.

Language	Articles	Content
Afar (Af)	Wikipedia	1
Muscogee (N-A)	Muscogee language, animals and plants in muscogee	2
Sichuan Yi (A)	man, rice, food, syllables, moon, light bulb, mountain, Liangshan, wheat, woman, Xichang	12
Hiri Motu (Aus)	The bible, John the apostle, a text excerpt	3
Marshallese (Aus)	Kahkun, hymn of the Marshall Islands, language of the Marshall Islands, Our father, Ave Maria, Marshall Islands, Book of Mormon, bread tree	8
Kuanyama (Af)	Days of the Week, Jehova's Witnesses, seasons, months, the bible	5
Choctaw (NA)	bear, goat, man, number, Human rights, language, colour, raccoon, Choctaw, Genesis, woman, Our father, Choctaw language	13
Ndonga (Af)	Ndonga, Duisburg, Uetersen, Ghana, Turkey, Wikipedia, Culture, Asia, Geography, Taiwan, the bible, 1st Epistle to Timothy, Gospel of Matthew, Pigazzano, Human Rights, Exodus, Nowy Dwór Królewski, English language, Genesis	20

Table 1: Articles in the very smallest Wikipedias as of March 2019. blue = traditional knowledge, green = story canons, red = infotainment, brown = language preservation, underlined = language (community) identity

languages. Only 4 Wikipedias lay below 95% (Kirundi, Latgalian, Ingush, Adyghe). Some articles featured no translation at all. On average these amounted to roughly 10% of articles, but Oromo was a slight outlier and constituted the maximum of 38%. In Oromo, a large number of articles which did not contain a link to a version in another language, referred to locally famous people hinting at a non-standard, probably need-driven usage of the Wikipedia. Overall, untranslated article contents were also describing local places, folk stories, local products, politics, entertainment related content and religious entities. 10,360 articles in the English Wikipedia were referenced in all, whereof 3,597 were referred to by more than one language, which is roughly a third suggesting some common interests between differing ICs. The top 10 articles referenced in most languages were Finland (36 languages), Turkey (35), Russia (34), China (33), Rio de Janeiro (33), United States (33), Italy (32), Norway (31), Chile (31), True Jesus Church (30). Countries, cities and languages are most referenced, then came some animals and plants, calendar month names, basic numbers, famous people like Confucius or Julius Caesar, scientific disciplines and religious entities.

In order to gain more insight into a categorization of these results, we extracted all English articles referenced and within them their English Wikipedia categories. Each category in a multiply referenced article was given a weight equal to the number of languages referencing that article. Additionally, a category tagged in more than one article increased its count multiply. The so-weighted most frequent Wikipedia categories referenced by the IC language articles can be seen in Table 2 in the leftmost column. With Living People on place 5, this category follows countries and places in obviously being of interest. The high incidence of categories starting in 'Member states' may in part be due to templates or automatic annotation. On rank 20, we find "spoken articles", consistent with our expectation that some ICs are still more rooted in orality or functional orality. On rank 25, we find the category tag *IUCN Red List least concern species* which points to an interest in the preservation of knowledge on biology and geography as attested above. Basic entities like Integers, Days of the year or Months may point to educational uses of the Wikipedia envisioned by many ICs.

Now, in order to mitigate the fact that Wikipedia categories can be very specialized, we mapped the Wikipedia categories to 44 WordNet categories condensed into the wordnet based categorization dictionary by the company Provalis.[5] We perfomed simple keyword matching on the category name and then counted the most frequent categories in the dataset. The top 10 categories can be seen in the second column of Table 2. Apart from corroborating the previous findings on locations and persons being of interest, we here see that artifacts and substances (relating loosely to geographical knowledge) also play a role.

In each language, we also annotated at least a random 10% of the articles for which (an English) translation was not available manually with the same categories after doing some research on the meaning of their content and came to a similar re-

[5] https://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/wordnet-based-categorization-dictionary/

Most referenced	WordNet categories (WNC)	WNC for untranslated
Member states (Ms) of the UN	Location	Cognition
Republics	Adjective	Location
Countries in Europe	Person	Communication
Ms of the Council of Europe	Pertainyms (abyssal, genetic etc.)	Group
Living people	Group	Animal
Ms of the Union for the Mediterranean	Communication	Person
Countries in Africa	Adverb	Plant
Ms of the African Union	Quantity	Food
English-speaking countries and territories	Substance	Artifact
Ms of the Organisation of Islamic Cooperation	Artifact	Artifact/Communication

Table 2: Categories of articles in descending order.

sult. The top 10 can be seen in the rightmost column of Table 2. The results do not differ much when looking at single languages instead of their union.

Finally, we investigated 3 more hints towards larger categories, namely the most frequent expressions in brackets in the titles of the English Wikipedia, which are often used to disambiguate and hence carry some domain information as for instance in 19_(number). The same was done for the categories, which themselves have bracketed expressions. We also extracted and counted all words in the categories before the genitive marker "of", Table 3 shows the results.

4 Discussion

While we found some topics common to the Wikipedias analyzed, here we seek to align these findings with the literature on the needs of ICs which could be answered by digital platforms.

1. ICs may need platforms to conserve traditional knowledge especially in the domains of local botany, zoology, kitchen and geography (Thomason, 2015, p.84), for instance, medical knowledge (Vandebroek et al., 2004), knowledge of local botany, knowledge on hyperaccummulator plants extracting metals from the soil in New Caledonia (Jaffré et al., 1976). This is reflected in various articles and their classifications in the Wikipedia.

2. There is a need for the preservation of the epics canon and oral literature of an IC, which Barrett & Cocq (2019) show to be useful in language preservation and education, helping to familiarize younger members of the communities with traditional concepts and worldviews as well as transmitting literary devices. These often circumspan newly acquired religous texts, which have been translated and since oral performance in sermons etc. is required often aligned with local cult. Again, we found Wikipedia articles with contents of entire stories.

3. An interest in entertainment or in broader terms infotainment is a need connecting humans from all language communities, for ICs consider Du et al. (2015, p.6). The use of social media to form larger global networks connecting with expats and other people may be located. Especially the description of local people and some non-standard usages of the Wikipedia pointed towards a reflection of this category. But also, the interest in countries and places may have some connection with this general need.

4. There is a wide consensus on the need for linguistic documentation of indigenous languages for various reasons, consider projects such as DoBeS.[6] Knowledge about smaller (missing-link) languages may significantly improve our capacities of understanding history, see Evans (2010, pp.138)[7] and of building accurate language family trees proposing new larger groupings like Na-Dene Yenisseian[8] (ibd. pp. 120).

Our empirical survey on Wikipedia articles and categories of their English translations found evidence for the needs identified in the literature, but the findings could be enhanced by two further points. One is trivial given the Wikipedia background, definitory information on basic entities. The other is something we would like to connect with the notion of language (community) identity (5.). Under this point a will/need to locate the IC within the cosmos of larger languages and communities seems (most) apparent to us, whilst it

[6] https://dobes.mpi.nl/

[7] The Udi language of the Caucasus is a peculiar example of a small language with an important history (Gippert and Schulze, 2007).

[8] Na-Dene as a family connecting Athapascan languages and Tlingit itself was recognized thanks to one single small indigenous language: Eyak.

In Brackets from Articles	In Brackets from Categories	Before 'of' in Categories
number	state	people
Cyrillic	U.S. state	descent
woreda (administrative unit)	country	rivers
genus	city	history
disambiguation	genus	order
city	music	languages
mythology	number	members
TV series	region	recipients
town	province	university
bird	Botswana	alumni

Table 3: Other indicators of categories referred to in the article titles.

crosses the 4 aforementioned needs. The preservation of the own language (4) is achieved/displayed by any activity in the digital of the IC, such as writing Wikipedia articles, but also by describing the ICs own languages. This raises the more general question of who writes the Wikipedia articles and who consumes them.

Some of the articles with very specialized content do at least suggest a very high familiarity with the respective cultures and at least a mild level of participation can hopefully be assumed in any of the cases examined as there are numerous examples from the literature where ICs do actively participate in preparing, creating, modifying and consuming internet content, compare for instance various articles in (Austin and Sallabank, 2011). Language learners and enthusiasts on the one hand, wiki managers on the other and computational activists may all play a part in the specifics of the single wikipedias. Much more in-depth studies[9] would be necessary to determine in how far, the results are warranted with respect to what they superficially appear to suggest. The problem with this is again the need for a familiarity with a vast number of sufficiently different languages which must at this point be conferred to future research. Whether pointing to the input of the language communities themselves or to language preservation activists or a mix of both, an interesting result is the across language similarity of the referenced entities or their categories (compare the numbers of multiply referenced English articles for instance).

A very tentative coloring schema has been applied to Table 1 above where blue refers to traditional knowledge (1), green to story canons (2), red to infotainment (3) and brown to language preservation (4). The underlined items are supposed to have meaning for language (community) identity. Although we are aware that this categorization is very tentative, by and large the presence of all of the categories, which have all been deduced from the literature is hopefully visible and extends also to the further categorization.

Unfortunaltely, we also found instances of 'predation' where sites of discussions in smaller languages were used for linking advertisment of completely unrelated content. Other articles were not in the minority language and/or concerned entities which were far from the lifeworld of the ICs, compare Table 1 and the Ndonga article on 'Nowy Dwór Królewski' which is an article that contains only the title plus two pictures where one is a shrine with an inscription in Polish. We assume these pages to have been written by bots or people experimenting with the Wikipedia (such as for automatic translation). However, luckily we found very few such instances.

5 Conclusion

We analyzed article titles in the Wikipedia for languages that featured less than 1,000 articles. Using various approaches of categorization (Wikipedia categories of the English articles, bracketing, WordNet etc.), we found some categories being commonly reflected, such as geo-locations, animals, plants and (famous) people. In the discussion, we surveyed the literature for apparent needs, a small potentially endangered languages community could have in the digital domain and tried to align these with our quantitative findings. The need to preserve the story canon and local traditional knowledge as well as infotainment are apparent from the topics of the articles which were largely similar in those languages analyzed. Ad-

[9]We would like to thank one of our reviewers for lining out some experiments with random pages or the close examination of single users edit histories, which we will take as an inspiration for further research.

ditionally, we interpret frequent categories such as countries as instances of a need to locate the IC within a larger global community. Knowing some of the needs those language communities could have towards the digital may enable better support. However, the results are to be taken with a grain of salt: it is not yet clear how much of the ICs efforts have really contributed to the results and how much of it has been the result of foreign contribution.

References

Antti Arppe, Jordan Lachler, Trond Trosterud, Lene Antonsen, and Sjur N Moshagen. 2016. Basic language resource kits for endangered languages: A case study of plains cree. In *CCURL 2016 - Collaboration and Computing for Under-Resourced Languages – Towards an Alliance for Digital Language Diversity (LREC 2016 Workshop), Portorož, Slovenia*, pages 1–8.

Peter K Austin and Julia Sallabank. 2011. *The Cambridge handbook of endangered languages*. Cambridge University Press.

James Barrett and Coppéline Cocq. 2019. Indigenous storytelling and language learning: Digital media as a vehicle for cultural transmission and language acquisition. In Coppéline Cocq and Kirk P.H. Sullivan, editors, *Perspectives on Indigenous Writings and Literacies*, number 37 in Studies in Writing, pages 89–112. Brill.

Rickard Domeij, Ola Karlsson, Sjur Moshagen, and Trond Trosterud. 2019. Enhancing information accessibility and digital literacy for minorities using language technology – the example of sámi and other national minority languages in sweden. In Coppéline Cocq and Kirk P.H. Sullivan, editors, *Perspectives on Indigenous Writings and Literacies*, number 37 in Studies in Writing, pages 113–140. Brill.

Jia Tina Du, Jelina Haines, Vicky Qiaoling Sun, Helen Partridge, and Dandan Ma. 2015. Understanding indigenous people's information practices and internet use: a ngarrindjeri perspective. In *Proceedings of the 19th Pacific Asia Conference on Information Systems (PACIS 2015)*. AIS Electronic Library.

Nicholas Evans. 2010. *Dying Words*. Wiley-Blackwell.

Joshua A Fishman. 1991. *Reversing language shift: Theoretical and empirical foundations of assistance to threatened languages*, volume 76. Multilingual matters.

Jost Gippert and Wolfgang Schulze. 2007. Some remarks on the caucasian albanian palimpsests. In *Iran and the Caucasus 11(2)*, pages 201–211. Brill.

Armin Hoenen. 2016. Wikipedia Titles As Noun Tag Predictors. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, LREC 2016.

Tanguy Jaffré, RR Brooks, J Lee, and RD Reeves. 1976. Sebertia acuminata: a hyperaccumulator of nickel from new caledonia. *Science*, 193(4253):579–580.

Steven Krauwer. 2003. The basic language resource kit (blark) as the first milestone for the language resources roadmap. *Proceedings of SPECOM 2003*, pages 8–15.

Walter J. Ong. 2013. *Orality and literacy*. Routledge.

Sarah G. Thomason. 2015. *Endangered Languages*. Cambridge University Press.

Ina Vandebroek, Jan-Bart Calewaert, Sabino Sanca, Lucio Semo, Patrick Van Damme, Luc Van Puyvelde, Norbert De Kimpe, et al. 2004. Use of medicinal plants and pharmaceuticals by indigenous communities in the bolivian andes and amazon. *Bulletin of the World Health Organization*, 82:243–250.

Mark Warschauer. 2000. Language, identity, and the internet. In *Race in cyberspace*, pages 151–170. Routledge New York.