A Snowball Sampling Approach for Studying Digital Minority Languages

Peter Nilsson
Swarthmore College
December 2014

Abstract

This thesis discusses the methods used to assess the linguistic diversity of the internet. I critique the current literature on internet language diversity, arguing that existing methods—which aggregate textual data from many languages, to the exclusion of video and audio data—are unsuited to the study of minority languages. To address these shortcomings, I propose a snowball sampling approach for studying an individual language's online use. I provide a series of case studies, in which I apply this method to several so-called "low-density" languages to demonstrate its potential. Finally, I conclude by describing how future studies could be improved through partial automation of the data collection process.

Table of contents

1 Introduction

The internet has the potential to serve as a powerful tool for language revitalization. It can provide a means for speakers of minority and endangered languages to circumvent geographical obstacles and form or maintain speech communities (Crystal 2000: 142; Grenoble & Whaley 2006: 190). In addition to allowing these *virtual language communities* (Holton 2011: 371), the internet can aid revitalization efforts by giving minority languages a way to advertise their presence and disseminate language-learning materials. However, both economic and technological barriers can prevent minority languages from being used on the internet, and the internet itself can serve to divide people along class and generational lines (Holton 2011: 372).

In order to make informed policy decisions about the internet, a better understanding of minority languages' online presence is needed. A number of authorities, such as UNESCO, have called for assessments of the internet's linguistic diversity (Paolillo et al. 2005). Accordingly, a range of studies have offered measurements of online linguistic diversity. Unfortunately, the methods that these studies use are only able to provide useful data concerning majority languages, as they either exclude minority languages from the data collected, or collapse data on all minority languages into the same category. As a consequence, most studies of the internet's linguistic diversity have only examined the relative distribution of the few most widely-used majority languages (e.g., Babel 1997; O'Neill et al. 2003; Wodak & Wright 2006). In this thesis I design a method to assess the online presence of individual minority languages, with the ultimate goal of revealing how the internet can work for or against language revitalization.

2 Terminology

Before discussing the literature, I will provide some terminology that will be useful for describing languages and the different ways in which they may be present online.

2.1 Digital minority language

In this paper I focus on languages which have a relatively small presence on the internet. These languages are often minority languages,[1] which, having a small population of speakers,[2] tend to have a small presence online. However, the degree to which a language is present on the internet does not always correlate with how widely it is spoken. For instance, the Nilo-Saharan language Kanuri has several million speakers and is not considered endangered—in fact, Kanuri enjoys regular use in public life, with radio and television programs produced in the language (Kornai 2013). Nonetheless, Kanuri has almost no detectable presence on the internet (Chew et al. 2011; Kornai 2013). Conversely, Inuktitut has a relatively robust web presence, ranging from government webpages written in Inuktitut to online Inuktitut games, despite having a much smaller population of around 30,000 speakers (Pasch 2008: 9).

As the quantity of language use on the internet does not always match the quantity of language use in the physical world, we need a term to identify languages that have little internet presence. The term with the widest use in the literature is "low-density language" (e.g., Karagol-Ayan 2007). This term is problematic because it implies that there is something intrinsic to the language that is lacking, thus attributing the absence of that language from online spaces to the

[1] I follow the definition given in the European Charter for Regional or Minority Languages, where a minority language is not an official language in the state in which it is spoken, and is spoken by only a subset of the total population (Council of Europe 1992).

[2] In this paper I only address spoken languages; sign languages are beyond the scope of this research. While I hope that the information I present is useful for the study of sign languages, it must be acknowledged that sign languages likely have a very different role on the internet, owing to the fact that they will have an almost exclusively video-based presence.

language itself rather than social or technological forces. Moreover, “low-density” suggests that instances of language use online are evenly distributed, and thus a language with little presence will be diluted by more commonly used languages. This does not reflect the reality that certain spaces exist where the use of a particular language may be very common, despite representing a small sliver of total language use on the internet. For instance, many Inuktitut webpages have a high density of Inuktitut content, with little or no content in other languages, despite the fact that Inuktitut webpages themselves represent a small portion of the internet. This density metaphor for language use might be appropriate for discussing a particular website where content in multiple languages is intermixed, such as a forum where Niuean speakers code-switch between English and Niuean (e.g., the OKA-KOA site described in Sperlich 2005), but it is not appropriate to discuss the use of a language on the entire internet in these terms. The idea of “density” is deceptive in that it suggests that low-density languages constitute a quantitatively-defined group, when in fact the division of languages into the categories of “high-”, “medium-”, and “low-density” is “of course… arbitrary” (Maxwell & Hughes 2006: 30), and how any particular author categorizes languages is seldom made explicit. Finally, the concept of a low-density language is problematic because it means that languages are categorized by how appealing a market they present to potential advertisers, rather than by any characteristic of the languages themselves.

Kornai (2013) treats language use online analogously to language use in the physical world. He refers to the absence of some languages from the internet as “digital language death.” This is an even worse label, as it implies that a language without online use has become obsolete, and ignores the fact that a language community could resurrect a language from its supposed digital death by simply starting to use the language on Facebook. While Kornai acknowledges

this possibility—terming it "digital ascent" (2013: 1)—the impermanence of a language's digital "death" is contrary to the sense of finality that "death" typically conveys.

I propose the term *digital minority language* to replace these problematic ways of discussing language use. This term has the advantage of using the analogy to minority language status in the physical world, thus acknowledging that these two statuses often co-occur, while making explicit the fact that we are discussing a phenomenon in the digital world rather than one the physical world. I define *digital minority language* as a language with a small enough quantity of online content that it could not be reliably located through random sampling. In this sense, a digital minority language is analogous to a hidden population (e.g., Heckathorn 2011: 356).

2.2 Language use versus language presence

At this point I would like to introduce a distinction between language *use* and language *presence*. *Language presence* is the broader of the two. Any online text, audio, or video containing content in a particular language constitutes an instance of that language being present on the internet. This does not require that any speakers of the language have used the internet, let alone put content on the internet in their native language. There are Pirahã words on the internet,[3] yet this is not evidence that people use Pirahã online. Data from linguistic research are sometimes available online, and thus provide an avenue for a language to gain an internet presence without online use by its speakers.

In contrast, actual *language use* on the internet entails that a speaker uses the internet to communicate with another speaker. There are a number of common edge-cases where it is not clear whether an instance of language presence should be counted as language use. One of these is the case where communication occurs in the physical world, but is recorded and uploaded to

[3] E.g., "Pirahã alphabet, pronunciation, and language," http://omniglot.com/writing/piraha.php

the internet. For instance, there are a number of videos in which a group of people talking amongst themselves in Anishinaabemowin happen to have been recorded by someone else.[4] The speakers may not have intended to communicate with anyone via the internet, but an internet user can listen to what they said. Does this constitute online language use? The answer will depend on the context in which the language is being studied, but, as Anishinaabemowin is a minority language, any study will likely focus on the extent to which the internet is useful for language revitalization. In that case, what is most relevant is whether other current or potential Anishinaabemowin speakers will watch the video.

This brings me to another division, that between unidirectional and multidirectional language use. *Unidirectional* use of a given language involves one party using the language without any accompanying response occurring in the same language. Most webpages represent unidirectional language use, because there is no mechanism for people who read the website to reply directly. For instance, the Nunavut Department of Health maintains a website that uses Inuktitut,[5] and thus Inuktitut speakers reading this site can passively engage with the language through reading. However, they cannot communicate directly with the speakers that created the content, nor with other speakers who find the same webpage; thus, the website does not afford them the opportunity to exercise their productive capacity for Inuktitut. In contrast, emails, instant messages, and tweets in one language can constitute *multidirectional* language use, because these forms of communication include a mechanism for response, and thus lend themselves to a conversation between people in which each participant can practice using the language.

[4] Examples include "Ricing Tools Part 1," uploaded to https://www.youtube.com/watch?v=6d5Rwu2Zfp8
[5] "Healthy Living," http://www.livehealthy.gov.nu.ca/iu

3 Review of literature

The current literature concerning online language presence is framed in terms of measuring the *linguistic diversity of the internet*. In other words, it is concerned with quantifying the number of languages that are online, and how much content they have. This broad approach is not intended to answer questions about specific languages: it studies languages in aggregate, rather than individually. In contrast, this thesis has a narrower goal—to provide a method for studying the online presence of any individual digital minority language. While there is no literature providing a general method for studying these languages, linguistic diversity studies provide a relevant background. I will overview this literature here, as the strengths and limitations of these studies can inform the design of a better approach.

There are a range of methods used to quantify the internet's linguistic diversity, and which of these methods is used in a particular study can strongly impact its results. Gerrand (2007) explains how these differences in methodology are responsible for such contradictions as those between studies claiming English use on the internet is declining rapidly and those claiming that English use is not changing. In doing so he provides a taxonomy for language diversity studies, which I summarize below.

Gerrand's taxonomy begins by describing the ideal metric for studying online language diversity, which he terms *user activity*. User activity encompasses all of the possible channels through which language can be used online, including publicly accessible data like webpages, tweets, and Facebook posts, as well as private data such as:

> email, Voice over IP, downloading software tools, playing multi-party computer games, downloading audio and video streams, etc. (Gerrand 2007: 1299)

Gerrand does not make the distinction that I have proposed between language use and language presence, and thus his definition of user activity does not explicitly exclude content in one

language which found its way onto the internet without the language's speakers purposely using the internet as a means of communication. Nonetheless, the spirit of his definition is analogous to my definition of online language use, in that it aims to describe the sum of all linguistic activity occurring on the internet. Crucially, Gerrand intends transient and private online communication to count as user activity, as he includes Voice over IP (VoIP) in this category. Thus video chatting through software such as Skype falls under this label.

Gerrand points out that there are no linguistic diversity studies that measure user activity over the entire internet, and there likely never will be. Much of what constitutes user activity is private, and studying private data presents both ethical problems and practical problems. Collecting private communications without users' consent is inappropriate, and even ignoring this ethical problem, would be difficult as it requires circumventing the systems that are designed to maintain people's privacy. As such, only surveillance organizations like the NSA have the capability to conduct such studies, and as Gerrand quips, "they have yet to publish their results" (2007: 1299). While there are a few studies that measure some of the components of user activity, those that do are limited in scope to a single forum (Wodak & Wright 2006), mailing list (Durham 2003), or collection of messages from newsgroups (Climent et al. 2003), and never fall into the same category as the linguistic diversity studies, whose scope is the entire internet.

For these reasons, surveys of internet language diversity have to resort to other metrics as a proxy for user activity. There are two main alternatives for conducting these studies, which Gerrand calls *user profile* and *web presence.*[6]

A user profile study estimates the number of internet users that speak a given language (Gerrand 2007: 1300). This method has the advantage of avoiding biases common in other

[6] Gerrand also identifies a third metric, the *diversity index*, which is a statistical measurement that uses an amalgam of different data to quantify language diversity. Few studies use this approach, and it does not differ enough from user profile or web presence measurements to avoid their flaws, so I do not address it here.

studies of online language diversity, such as the relative ease of finding language content that is written rather than in video or audio form. However, this metric is very far removed from actual user activity. In fact, no data needs to be collected from the internet for this type of study. Gerrand discusses how one of the authorities on internet language presence—the marketing company Global Reach—calculates user profile. Global Reach simply multiplies the number of speakers in each country by the proportion of people in that country with internet access, and then assumes that this estimate of how many people could potentially use a given language on the internet reflects how many people actually do. Gerrand observes that this ignores several key facts. Firstly, speakers of different languages within the same country may not be equally likely to have internet access. Secondly, multilingual speakers may use a different language on the internet than their native language. Finally, the population data, language usage data, and internet coverage data for a given country are often measured in different years, and are often years out of date; therefore, multiplying together these disconnected numbers may produce wildly inaccurate results. Even if these estimates are close enough for marketers to target their advertisements in a profitable fashion, they are useless to any linguistic study of how language is actually used online.

The other major category in Gerrand's taxonomy takes a more reasonable approach. *Web presence* studies measure the number of publicly available webpages in a given language. As existing studies aim for a statistical understanding of language diversity on the internet, they require large amounts of content from webpages which can be automatically analyzed by a computer. This means they rely on software to be able to recognize what language is used for each datum. This further restricts the scope of these studies to written languages that have software support in the form of fonts for their orthographies. Video and audio data, even from

languages for which language detection software exists, cannot be used, which is a major flaw in this type of study. Furthermore, in surveying only publicly accessible webpages, these studies neglect the potential role of social media in language use.

While the studies reviewed by Gerrand vary in the methods they use to quantify language diversity, most share a major limitation: they depend on software that can only identify the most prominent languages on the internet. This may be sufficient to answer questions about the relative use of English in the online world, but is of little use for research on digital minority languages. The studies which Gerrand describes almost exclusively examine the presence of majority languages. The two studies Gerrand highlights as providing the most comprehensive information on minority languages were both conducted over ten years ago (Guinovart 2003 and Mas i Hernàndez 2003, cited in Gerrand 2007). These both used the search engine AllTheWeb, which enabled searching in 48 languages. While a few of these languages were minority languages, these were all European; in fact, only a few languages indexed by AllTheWeb—such as Chinese and Japanese—were not from the Indo-European family. This hardly constitutes a representative sampling frame for majority languages, let alone minority ones. Not only is this method impossible to extend to other minority languages, the original studies themselves are no longer replicable as the search engine AllTheWeb no longer exists (Gerrand 2007: 1309).

The limitations of these studies make them unsuited to finding minority language content. Even if minority language content was indexed by search engines and thus possible to capture in a similar study, the amount of minority language content is low enough as to not be informative when folded into a large collection of data. On the scale of the entire internet, digital minority language content is not common enough to show up in any percentage-wise breakdown.[7] Little

[7] For instance, Gerrand (2007) overviews the results of a study that sampled 30,000 potential IP addresses (Babel 1997). It gives a percentage-wise breakdown of the fifteen most commonly-used languages in the sample. None

would be gained even if this breakdown had a finer resolution—there isn't much that can be inferred about *how* a minority language is being used by looking at a single number.

Due to these methodological limitations, there has been a paucity of studies examining the online presence of digital minority languages. A recent study by Kornai (2013) has filled this void, and thus has become the de facto source for claims about language diversity on the internet, yielding such headlines as "How the internet is killing the world's languages" (Dewey 2013). Kornai adapts the EGIDS[8] rankings to the internet, aiming to create an analogous scale that measures language endangerment in the online world. His new scale distinguishes four categories, which he terms *Thriving*, *Vital*, *Heritage*, and *Still* (Kornai 2013: 1). The distinction between Thriving and Vital is unclear, and not simply one of quantity—while Kornai mentions that the two could be distinguished by the number of language users, he explicitly rejects this as arbitrary. The distinction between these two categories and Heritage is clearer—a Heritage language is present on the internet, but due to an effort at preservation rather than everyday use. The final category of Still applies to languages without any presence on the internet.

This categorization is problematic. EGIDS rankings take into account which generations with a community use a language—the distinction between labels 6a-b, 7, and 8a-b is that of whether children speak the language, only parents and grandparents speak the language, or only grandparents speak the language (Lewis et al. 2014). In collapsing the 13 EGIDS categories to a set of four, Kornai ignores this set of distinctions, assuming that "once some speakers transition to the digital realm, their children and grandchildren automatically do so" (Kornai 2013). While it is reasonable to assume that a generation which uses the internet will pass on the tradition of internet use to its children, it does not necessarily follow that the children will use the same

---

were minority languages; all remaining possible languages (5.6% of the data) are lumped into the "None or unknown" language category.

[8] Expanded Graded Intergenerational Disruption Scale, see Lewis et al. (2014)

language on the internet as their parents. Thus, even if a language has a strong presence on the internet due to a very technologically-inclined generation of language users, this may change when it comes time for the next generation to choose how to interact with the online world.

Kornai also argues that the EGIDS distinction between international, national, and regional languages is inappropriate for assessing use on the internet—which transcends national boundaries—and thus discards it. While this is reasonable, Kornai does not replace this distinction with one more appropriate for the internet. He thus ignores any distinctions that exist in the type or size of language communities. For instance, it might be useful to distinguish between a language that is spoken only within one geographical region, and one that is spoken by a diaspora with no main geographical region. Such differences could affect the prognosis of the language's use on the internet, as well as the assessment of how important it is that the language be used on the internet—a language community that is physically separated will rely more on the internet to keep in contact than a language community where speakers can interact directly.

More importantly, the analogy between language endangerment and language use on the internet is itself flawed. EGIDS can make the assumption that a language that goes extinct will not come back, because native speakers of a language must grow up in an environment where the language is spoken regularly. If there are no speakers left, new speakers will not spontaneously appear. However, language use in the digital world is not restricted in this way. A language that is not (yet) used on the internet, or that has ceased to be used, is not necessarily doomed to a "digital language death" in Kornai's (2013: 1) terms. Rather, the status of a language outside of the digital world is what determines its future online. If the reason for a language's absence from the internet is the extinction of the actual speech community, then it will indeed continue to be

absent from the internet. In contrast, as long as a language is spoken, it retains the potential to be used on the internet in the future. It is therefore misleading to speak of language death on the internet as a phenomenon separate from language death in the non-digital world. Furthermore, the idea of digital language death serves to conflate the lack of internet presence with language death, which is what allows Kornai to conclude that there is a "massive die-off of the world's languages," when in fact he merely found evidence that few languages are used online.

These shortcomings leave the issue of how to measure digital minority language use on the internet an open question. In the remainder of this paper, I present a new approach that is designed with digital minority language research in mind.

4 Sampling digital minority language content

I have argued that the current literature on internet linguistic diversity neglects digital minority languages. This is in part due to the reliance of linguistic diversity studies on random sampling, which is necessary to produce estimates of language presence that are comparable between languages. I contend that the status of digital minority languages is best studied individually, one language at a time. I therefore reject the question of how many languages are on the internet as unproductive. My goal is to develop a method that is suited to answering questions about the use of a particular language on the internet, rather than questions about language use at large. The following are examples of questions which fall within the scope of this method:

- How strong a presence does language X have on the internet?
- Through what media (video, audio, text…) do people communicate in this language?
- Does the presence of the language online constitute actual language use?

- For what purpose is this language used on the internet? Are fluent speakers using it for day-to-day communication, or are most users non-speakers trying to learn the language for the first time?
- What variety of the language is used online?
- What obstacles are there to using this language on the internet?

These are all questions that cannot be answered through the random sampling approach used in traditional linguistic diversity studies. In the rest of this paper, I discuss how an alternative approach might be implemented.

4.1 Choosing a sampling method

Snowball sampling commonly refers to a nonprobability sampling technique used in sociology to study hard-to-reach populations (Heckathorn 2011: 356).[9] Snowball sampling involves starting with an initial seed made up of people from the population being studied. These people are then asked to recruit more research participants from the same population. This process is repeated, with participants continuing to recruit other population members; the sample thus "snowballs" to an increasing size. In this way, snowball sampling can be used to access members of a population that could not feasibly be located by random sampling.

Digital minority language content cannot be located in substantial quantities by random sampling—for instance, Gerrand (2007) points out that a previous study sampled only a few times as many webpages as there are written languages, and thus had very little chance of retrieving any datum from a digital minority language (Lavoie & O'Neill 1999 and O'Neill et al. 2003, cited in Gerrand 2007). I suggest an analogy between hard-to-reach populations and digital

[9] This differs from the original use of the term s*nowball sampling*, which was intended "as a means for studying the structure of social networks" independent of whether a nonprobability approach was needed (Heckathorn 2011: 356).

content that is hard to find. In this analogy, instances of digital minority language content that have been located are like respondents in a sociological study. Following the analogy to sociological research, digital content is connected not by social networks but by hyperlinks. We can thus select an initial set of seeds—perhaps by using a search engine—and snowball out by following hyperlinks in the content we have found. This approach might be more precisely described as *dirty snowball sampling*;[10] while human respondents in a snowball sampling study will generally recruit other members from the same populations, hyperlinks are much less reliable, as they often lead to content in a different language. Thus, removing irrelevant data from the snowball is more crucial to the process I describe than it typically is for sociological studies.

The dirty snowball approach stands in contrast to that of the linguistic diversity studies discussed previously, which typically use random sampling (e.g., Lavoie & O'Neill 1993; Babel 1997; O'Neill et al. 2003), or search engine results alone (e.g., Mas i Hernàndez 2003; Guinovart 2003). A study opting to use snowball sampling sacrifices the ability to make quantitative conclusions, as it does not approximate a representative sample. However, it gains the ability locate large quantities of digital minority language content. This gives snowball sampling studies the capability to examine the use of endangered and minority languages online, filling a major gap in the literature. Furthermore, the data that snowball sampling provides can inform later quantitative studies; thus, having this method available is beneficial even for the quantitative study of online linguistic diversity. Finally, the greatest advantage of dirty snowball sampling lies in its similarity to real-world internet use. It is a naturalistic way of sampling online content, because it is essentially what people do when they use the internet—use search engines and

[10] I owe this term to Eric Nilsson, who suggested that it better summed up the problem of encountering pollution from other languages in the sample.

hyperlinks to locate new content—and therefore, any biases introduced by this sampling method will be similar to those experienced by internet users. Hence the content that this method locates will be fairly representative of what people would actually find. Thus the main "flaw" of snowball sampling is actually a benefit here—biases in the collection of data are informative, because they reflect what content internet users are likely to find.

4.2 Procedure for manual sampling

Harvesting data from the internet is typically an automated task, performed by a computer. The sampling approach which I have described could ultimately be automated as well. However, only a partial degree of automation can be achieved—because this method is to be employed for studying digital minority languages, there will rarely if ever be software that can identify the language of interest. Furthermore, much valuable content might be stored in video or audio form, and thus could not be read by a text-based language detector even if one existed for the language in question. As a consequence, the parts of the sampling process that require language identification will always require some human intervention.

In a later section (§6), I will describe how those tasks that can be done programmatically could be automated. Before I address how to implement an automated version of this method, I will first discuss how a researcher might search for a digital minority language manually. The steps required are outlined below:

1. *Initial research*. The researcher gathers information that is already known about the language being studied (henceforth the *target language*). The most important data are names: the name of the target language in English, the name of the target language in other languages spoken by the same group of people, and autonyms for the language. If a lexicon is available, it may be helpful to choose some common vocabulary items in the

target language that are unique enough to not be found in other languages.[11] This initial research should be used to create a set of strings that are likely to return relevant results when supplied to search engines.

2. *Finding the first pieces of content*. Unless the language in question is unlucky enough to share a name with some more common search term, search queries including its name are likely to return some content *about* the language—however, the content may not be *in* the language. This is where knowing some words unique to the language may yield better results. Failing that, digging through enough pages of search results for the language's name will likely yield content eventually. This may require using several different search engines, potentially with settings for different countries or languages. Searching on Facebook or YouTube might succeed even when searching on Yahoo and Google has failed.
3. *Identifying what content is in the correct language*. Once the researcher has found some relevant content, it is necessary to determine what language the content is in. For this reason, it is important that the researcher either be able to understand the target language, or work with a consultant who has this expertise.
4. *Snowballing*. After locating some relevant content that is in the correct language, more content may lie behind the hyperlinks on the page, or in the case of a site like YouTube, in related content that is "suggested" by the website. In many cases it is useful to follow links even from pages that are not in the correct language; for instance, some Anishinaabemowin sites are mostly in English, but have links to pages with lots of Anishinaabemowin.

[11] In some cases, these lexical data may be more useful than autonyms. Some language may have a very short name—e.g. *Ho*—or share a name with other languages (K. David Harrison, personal communication December 16, 2014), and so their names may have little potential to locate the correct content.

Having described a procedure for locating digital minority language content, I now present examples of the procedure in use.

5 Case studies

I conducted a series of case studies to assess the reliability of this snowball sampling method, and to further refine it. For each case study, I selected a language that I knew or inferred had a small internet presence, and used the method I have outlined to locate as much content as possible. I have presented a summary of the results here.

These case studies are a proof-of-concept, but do not represent the full potential of this sampling method. In a real use-case, research on any language would be conducted by experts who have at least a basic ability to read and speak the language. I do not have this level of expertise in any of the languages I examined, so I may have missed some opportunities to collect data, or, conversely, may have incorrectly gathered some data from languages closely related to those I intended to study. These do not represent limitations of the method I have described, but rather, limitations in what can be accomplished by a non-expert in a short timeframe.

While conducting these short case studies, I found substantial variation in the kind of online presence that each language I examined had. The languages I searched for ranged from having barely any online presence at all, to seeing regular online use supported by government policies and aided by technological support. Although an exhaustive taxonomy of the patterns of online language presence is beyond the scope of this paper, I incorporate a tentative categorization of the modes of language presence that I have observed into my summary of these case studies. The categories I propose are presented in §5.1.

5.1 Categorization of digital minority language presence types

Coincidental presence

This category covers all online occurrences of a language that were not intentionally produced for the internet. This includes television broadcasts that were recorded and later uploaded, videos of everyday speech in which the speaker was not intending to communicate with an online audience, and old grammars of the language that have been scanned and archived online. The common theme for this category is that coincidental content represents neither online use of the language, nor a conscious effort to create content for online use.

Descriptive presence

Linguistic descriptions of a language constitute descriptive presence. Descriptions can cover any of the components of the language—its phonology, its morphology, its lexicon, and so forth. A language with a descriptive presence has some scholarly work concerning its structure publicly available online. However, I do not include old grammars that have been archived in this category, because I intend the idea of descriptive presence to capture intentional online language presence, and books that were originally intended to be distributed in hard copy do not follow this pattern.

Educational presence

Content that is intended to teach people to speak the minority language in question falls into this category. In some instances educational presence is easy to distinguish from descriptive presence; at other times, the two are harder to separate. In both cases, language content is being presented with the intent of instructing some audience. The choice of audience differentiates the two—descriptive presence occurs when linguists are the main audience, and educational

presence occurs when the main audience is potential speakers. Holton (2011: 374) makes a similar distinction between *archival* or *preservation* formats of endangered language content on the one hand, and *presentation* formats on the other.

Utility presence

This category represents the clearest indication that a language has a robust internet presence. In cases of utility presence, people use a minority language online not to teach or practice it, but because it is useful for some other purpose. This can only occur when enough people use the internet to make it practical for non-linguistic goals: if speakers can't rely on other users to understand their language, they will resort to a more widely spoken alternative. Of the languages I examined, only one had a clear utility presence.

5.2 Languages with a minimal presence

This category represents languages at the extreme low end of internet saturation. They have some small presence on the internet, but this is due primarily to actors outside of the community, such as linguists or missionaries. These languages do not see any online use in the terms that I have defined it, as no speakers use the internet to communicate in their language.

Yokoim: A language with a small descriptive presence

Yokoim, also known as Karawari or Tabriak (Booth 2014), is spoken in the East Sepik province of Papua New Guinea by about two thousand people (Anderson & Harrison 2014; Lewis et al. 2014). Although Yokoim speakers may not have internet access, some Yokoim has recently been uploaded to the internet: a talking dictionary[12] and a YouTube video of a song in Yokoim sung

[12] "Yokoim Talking Dictionary," http://talkingdictionary.swarthmore.edu/yokoim/ (Anderson. & Harrison 2014)

by a native speaker, with an English translation.[13] This content was uploaded by linguists under the direction of the Yokoim community; while Yokoim speakers are not currently using the internet (K. David Harrison, personal communication November 14, 2014), they nonetheless have intentionally created a Yokoim presence online.

I classify Yokoim's internet presence as descriptive. While there is currently little Yokoim content online, the little that exists is very well annotated—the talking dictionary includes phonemic transcriptions paired with audio and English glosses, and the song on YouTube is accompanied by an English translation. Although this content could be used to teach Yokoim to new speakers, the main audience currently consists of linguists. Yokoim clearly does not have multidirectional use on the internet; this is impossible without Yokoim speakers being online. Whether it has unidirectional use is less clear. The Yokoim content that is currently online was produced by speakers with the intention of having it available on the internet; however, the content reached the internet by proxy, in that it was uploaded by linguists.

One of the characteristics distinctive of a language with a strong descriptive presence is the ease with which it can be found online. While Yokoim content occupies only a minute fraction of the entire internet, it is trivial to find this content with a search engine. This is due to the fact that the little Yokoim content that does exist is clearly tagged as being in Yokoim, as it is presented in the context of a linguistic discussion. This makes Yokoim disproportionately easy to locate, considering its small body of content.[14]

[13] "'Imba Us', a song by Louis Kolisi in Yokoim (PNG)," https://www.youtube.com/watch?v=OK6dp4QOplw

[14] In contrast, not all communities have a name for their language (Ted Fernald, personal communication December 16, 2014). This makes them much harder to locate through this method.

Kanuri: A language with a small and mainly coincidental presence

Kanuri is a group of Nilo-Saharan languages with around 3.7 million speakers (Lewis et al. 2014) located primarily in Nigeria and Niger (Löhr et al. 2009). Kanuri was identified as having a very small internet presence in Kornai (2013), despite the fact that it has a relatively large population of native speakers and is not endangered in that at least Central Kanuri has a low EGIDS ranking of 3 (Lewis et al. 2014). It is interesting to note that Chew et al. (2011) failed to find Kanuri online at all. Though it is possible that Chew et al. were simply less thorough, this discrepancy may be due to a difference in their methods for locating languages. Both studies located languages by drawing from a set of websites that contain content organized into a large number of different subdivisions for various languages (such as Wikipedia), so the difference in results might be due to the set of sites they used.

The discrepancy in results suggests that Kanuri is present on the internet at just the threshold of detectability. Although Yokoim also has a minute internet presence, it presented a less thorough test of my search method because of how clearly identified it is, with all content accompanied by English descriptions of the language. In contrast, Kanuri content was hard to find. While it is possible that the difficulty in detecting Kanuri is primarily the result of there being little content, my search revealed a number of important factors that previous authors did not address. In particular, there are some extra-linguistic reasons that complicate the search.

It is difficult to find search results for Kanuri in part because so many of the results are for unrelated things with similar names. Kanuri has had bad luck in this regard: *Kanuri* appears to be a common surname, so many of the Google hits for this string are simply people. *Manga*, the name of one of the dialects of Kanuri, mostly yields results relating to the Japanese comics, rather than the language. *Beriberi,* another word that sometimes used for Kanuri, returns results

for a disease of the same name caused by thiamin deficiency. *Tumari,* the name of another Kanuri dialect, yields some correct results interspersed with many results for the Indian television show *Tumhari Paakhi* ('Your Paakhi').

Using Yahoo instead of Google yielded similar results. Somewhat more relevant results came from using the Google's pages for Nigeria[15] and Niger.[16] Searching for results in French yielded essentially the same results as searching for results from the page for Niger, which is by default set to search in French. This indicates that the language used for searching—even when it is not the target language—can impact the results.

Kanuri can be written in an Arabic script as well as a Latin one. However, I have very little familiarity with Arabic, so I cannot tell which results are in Arabic rather than Kanuri, although I suspect most results in this script are in fact in Arabic. This is a different situation from Anishinaabemowin, which is written in a Latin script, and occurs in the same online environment as a language I am familiar with (English). This represents a limitation in my own knowledge, rather than a limitation inherent to this method of snowball sampling—in practice, a linguist with some knowledge of Arabic, or with a collaborator who could read Arabic, would be performing the search.

The small amount of content that does appear to be Kanuri came in the form of YouTube videos. This included a recording of a Kanuri television show,[17] as well as a full movie.[18] It is unclear to what extent the movie is a good example of Kanuri, as it is posted on what appears to be a Christian missionary channel, and, like the other movies posted by that channel, it is an originally English movie dubbed into a new language. Thus I classify Kanuri's internet presence

[15] www.google.ng
[16] www.google.ne
[17] "Kanuri Drama," https://www.youtube.com/watch?v=ju7oB5wWT9E
[18] "The Story of Jesus for Children," https://www.youtube.com/watch?v=iGG7yG1TUcU

as mainly *coincidental*, as the videos in Kanuri were not uploaded in the interest of communicating or increasing Kanuri's internet presence, and may not have even been uploaded by Kanuri speakers.

5.3 Languages with some use

Niuean: Occasional online use by a diaspora

Niuean is an Austronesian language with around ten thousand speakers, though estimates vary (e.g., Lewis et al. 2014, Sperlich 2005). Niue, the small island nation where the language originated, now has only one or two thousand residents; the population is declining as Niueans emigrate to New Zealand (Sperlich 2005). Although Niuean's status as an official language of Niue garners it an EGIDS ranking of 1 (Lewis et al. 2014), Niuean is recognized by UNESCO as "definitely endangered" (Mosely 2010), and it is being supplanted by English, even in local radio and TV ("The language context of Pacific countries" 2005; Sperlich 2005).

Sperlich (2005: 51) investigated the use of Niuean on the internet, asking the question, "Will cyberforums save endangered languages?" As the Niuean speech community is now divided between two physically distant locations, the internet provides a unique means for Niuean speakers to overcome their distance. Sperlich found that the OKA-KOA forum[19] did feature people corresponding in written Niuean. Unfortunately, this website appears to be gone—it has been down for "maintenance" since the Wayback Machine[20] first archived it in January of 2011. My own search for Niuean did not produce any websites with written Niuean conversations. However, I found a number of videos where Niuean is spoken, as well as several Facebook pages that have occasional exchanges in Niuean, interspersed in the mainly English

[19] "TalaNet Niue," talanet.okakoa.com
[20] "TalaNet Niue," https://web.archive.org/web/20110601000000*/http://talanet.okakoa.com/

conversations (Fig. 1). There is one website ("Learn Niue") with a few Niuean phrases that have both English translations and audio recordings.[21] Additionally, the website of Radio New Zealand International has a collection of short broadcasts in various languages, including Niuean.[22] This listing appears to be updated almost daily.

Figure 1. Screenshot from the "Niue Bling" public Facebook group, showing a post from the summer of 2014. Some Niuean text is interspersed with a greater volume of English text. This group can be found at https://www.facebook.com/groups/760345027338759/.

While the position of Niuean on the internet may not be as strong as it was during Sperlich's study, it is nonetheless clear that Niuean is actually used online. This stands in contrast to Yokoim and Kanuri, whose presence does not constitute clear online use. Some of this use is multidirectional, as evidenced by conversations on Niuean Facebook groups. There

[21] "Learn Niue," http://www.learnniue.co.nz/learnniueanlanguage/
[22] "News in Pacific Languages," http://www.radionz.co.nz/international/programmes/pacificlangaugesnews

does not appear to be any website dedicated to explaining the grammar of Niuean to an audience of linguists, so Niuean does not have a descriptive presence on the internet. It does, however, have a small educational presence in the form of the "Learn Niue" website. The radio broadcasts in Niuean constitute a form of utility presence.

Anishinaabemowin: Some online use with a focus on revitalization

Anishinaabemowin, also known as Ojibwe, is a group of several closely related dialects spoken in North America (Noori 2011) – or in the Ethnologue's terms, a macrolanguage (Lewis et al. 2014). There are slightly over 90 thousand speakers by the Ethnologue's count. The language is dying, as most speakers are over sixty years old, and "no one learns Anishinaabemowin as a first or only language anymore" (Noori 2011: 3). Despite these facts, or perhaps because of them, Anishinaabemowin is known to have an internet presence: Noori created one website (Noongwa e-Anishinaabemjig: People Who Speak Anishinaabemowin Today) as a teaching aid for the language.

Locating Anishinaabemowin content was straightforward. I began by searching through Facebook. I found that there are at least three relevant Facebook pages, along with one Facebook group dedicated to Anishinaabemowin. As a number of posts referenced videos with Anishinaabemowin, I proceeded to YouTube. There are at least eight YouTube channels with content in the language. Some of the videos consist only of an Anishinaabemowin vocabulary item with an English explanation, whereas others involve conversations among several people in Anishinaabemowin.

Outside of social media, there are more than ten websites with Anishinaabemowin content. These vary in the amount of content that is in the language, but no website consists only of Anishinaabemowin. This is in part because, as there are no new native speakers, websites

focus on teaching younger English speakers Anishinaabemowin. As a result, the navigational links and most of the text consists of English.

While a large amount of Anishinaabemowin vocabulary is used online, the general rule is that people are talking to each other in English about Anishinaabemowin. There is little communication directly in the language, particularly in written form. This is why despite the plethora of language-learning resources, there is no Anishinaabemowin Wikipedia, and the Twitter users identified by Indigenous Tweets (Scannell 2014) are in fact primarily using English, with some Anishinaabemowin words. Nonetheless, it is clear that Anishinaabemowin sees more real use online than do Yokoim, Kanuri, or Niuean. Overall, Anishinaabemowin has an educational presence, as virtually all of its content is occurs in the context of language learning. While much Anishinaabemowin use occurs in the unidirectional context of language-learning resources, some use occurs multidirectionally on Facebook.

5.4 Languages with regular use

Inuktitut: Regular use with official support

Inuktitut is an Inuit language spoken by around 30,000 people in Canada (Pasch 2008). Like Anishinaabemowin to the south, Inuktitut is a minority language whose use has declined in recent times (Pasch 2008). It too has a fairly strong internet presence. Unlike Anishinaabemowin, Inuktitut is frequently used on the internet as a medium of communication in and of itself (utility presence), whereas Anishinaabemowin's use is largely pedagogical, geared towards teaching non-speakers the language. Even in digital spaces geared towards the teaching of Inuktitut, there is a clear contrast between the degree of use of each language. For instance, there are a number of public Inuktitut Facebook groups for language teaching, just as for there are for Anishinaabemowin. However, while posts on Anishinaabemowin pages tend to consist of

English with some Anishinaabemowin vocabulary,[23] posts on the Inuktitut pages are often largely or only in Inuktitut (Fig. 2).

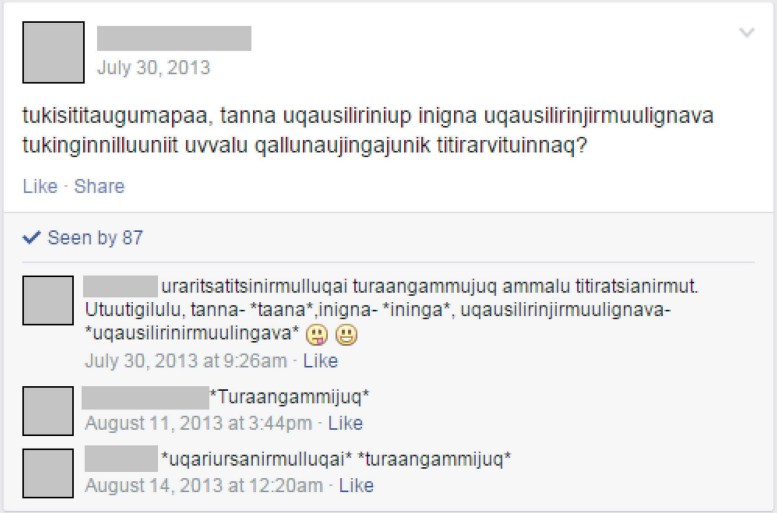

Figure 2. A post on an Inuktitut Facebook group, demonstrating the use of the language. See https://www.facebook.com/groups/105727982918169/.

Another major difference that sets Inuktitut aside is its degree of official use. There are Canadian government websites that offer content in Inuktitut as well as English and French. Inuktitut text is sometimes written in the Latin alphabet and other times is written in Inuktitut syllabics. In this sense Inuktitut faces a challenge that Anishinaabemowin does not—software support is necessary to view webpages written in the syllabic script. However, this support is easy to obtain; computers may come with support for these characters, and for those that don't, the Canadian government offers a download.[24] While using a separate character system can cause difficulties, the payoff for websites using syllabics is a distinctly Inuktitut impression.

[23] Similarly, Holton (2011: 391) observes that "blogs for indigenous American languages tend to provide information about indigenous languages through the medium of English."
[24] Available at "Pitquhiliqiyikkut," http://www.ch.gov.nu.ca/in/ComputerTools.aspx

Indicative of the strength of Inuktitut's web presence is extent to which entities that are not legally obligated to provide Inuktitut content cater to the language community. For instance, Greenpeace has a portion of their website in Inuktitut,[25] there is an Inuktitut Bible translation,[26] and there is at least one translation service that offers to render text into Inuktitut.[27] These all constitute a utility presence for Inuktitut.

One of the most impressive features of Inuktitut's online presence is a website that offers a number of small Flash and Javascript games aimed at teaching Inuktitut. There are also several games that can be downloaded to a desktop or iPad (similar to some offered in Anishinaabemowin). These are all hosted on the Katavik school board's website.[28]

Inuktitut thus enjoys regular use online, both unidirectional and multidirectional. It has content constituting both educational and utility presence. A portion of this use is due to official recognition of language and subsequent use on government webpages. Inuktitut would likely have not had as strong a presence without the websites and technological support provided by the Nunavut government.

5.5 Summary of case studies

The results from all five languages are summarized in Table 1. Yokoim and Kanuri differ from the other languages in only having only one kind of web presence and little if any online use. Category-wise, the differences between Niuean, Anishinaabemowin, and Inuktitut are less stark, although Inuktitut clearly has the strongest presence. I tentatively mark Yokoim as having some

[25] "ᐅᑭᐅᖅᑕᖅᑐᖅ," http://www.greenpeace.org/canada/en/campaigns/Energy/Arctic-Inuktitut/
[26] "1 ᒨᓯᓯᑦ 1 ᐊᒡᓛᑦ ᐃᑦᔪᖅᖕᓃᑦᑐᑦ," https://www.bible.com/bible/455/gen.1.eaib
[27] "Professional Inuktitut translation service," http://www.tomedes.com/inuktitut-translation.php
[28] "ᐃᓄᒃᑎᑐᑦ ᖃᕆᑕᐅᔭᒃᑯᑦ ᐱᙳᐊᕈᑏᑦ," http://www.kativik.qc.ca/iu/inuktitut-qaritaujakkut-pingnguarutiit; also available in French and English

unidirectional internet use, but concede that the indirect route by which Yokoim entered the internet may mean it is inappropriate to consider the community to be using its language online.

Table 1. Categorization of internet presence and use by language.

Language	Presence				Use	
	Coincidental	**Descriptive**	**Educational**	**Utility**	**Unidirectional**	**Multidirectional**
Yokoim		✓			(✓)	
Kanuri	✓					
Niuean			✓		✓	✓
Anishinaa-bemowin			✓		✓	✓
Inuktitut			✓	✓	✓	✓

6 Automating the sampling method

I suggested previously that some of the sampling process could be performed by an automated tool. What work could this hypothetical tool take over from a human researcher? Menial tasks that lend themselves to automation include querying search engines for data, discarding results that are obviously in the wrong language, recording content that has already been examined, and keeping track of links that should be followed in the next wave of snowball sampling.

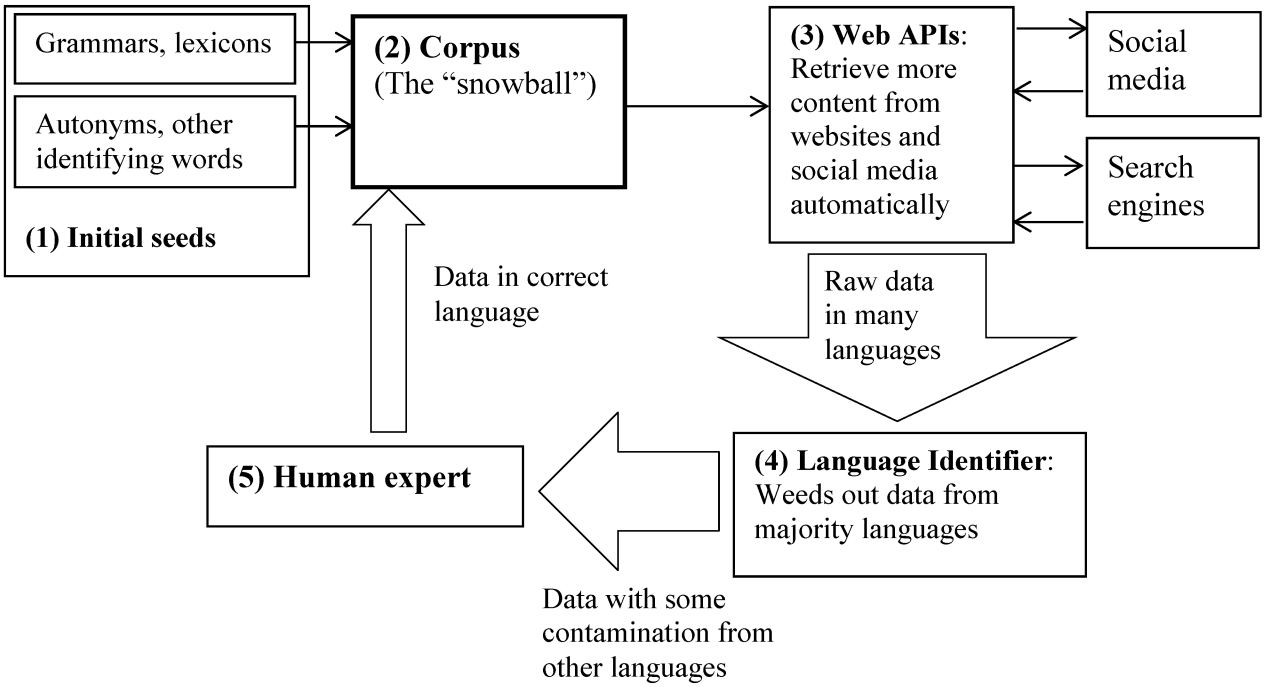

Figure 3. Schematic representation of how the sampling could be (partially) automated.

I present a possible workflow that incorporates both automated and non-automated components, diagrammed in Fig. 3 above. I discuss the first (1) step in this procedure—the choice of initial seeds—in §6.1. Steps (2-5) are repeated for each wave of snowball sampling; they are detailed in §6.2-6.5 below.

6.1 Gathering the snowball's initial seeds

The first stage of the partially-automated snowball process is one of the two that is not automated. Researchers must start the snowball rolling by finding search terms that are likely to locate data in the target language. Good candidates for search terms include autonyms, words listed in lexicons and grammars that may be unique to the language, and any other words that

could identify the target language, the region it is spoken in, or the people that speak it. Once enough identifying words have been chosen, the researchers must use search engines to locate a few pieces of content which use the target language. This is essentially the same process that was described earlier under the manual sampling method (§4.2).

6.2 Storing the corpus

The initial seeds gathered in step (1) are stored in the corpus (2), which is implemented as some electronic database. The exact contents of the corpus may vary based on the needs of the researcher, but each entry in the corpus will likely contain content in the target language, the URL from which the content was retrieved, and various metadata. This corpus also constitutes the "snowball" of the snowball sampling procedure. Therefore, corpus entries will all store a list of references to other potential data. These references include hyperlinks to be followed, names of people or places that might be unique enough to yield more data in search engine queries, and—for data gathered from social media—user accounts that are associated with the data. The corpus will keep a record of which of the leads to new data have already been followed, so that the same content is not repeatedly gathered into the snowball.

6.3 Retrieving more data with web APIs

Many social media websites and search engines offer application programming interfaces (APIs), which specify procedures for automatically requesting content.[29] In step three (3), the snowball sampling tool runs through the list of hyperlinks and other potential leads that are stored in the corpus. It gathers new content by following all hyperlinks and by requesting new information from search engines and social networking sites via their APIs. It then records the sources as

[29] Examples include the Facebook Graph API, Twitter REST APIs, and Yahoo BOSS Search API (Facebook 2014; Twitter 2014; Yahoo 2014).

used in corpus, so that future waves of snowball sampling don't request this data again. Automating this step saves researchers from the tedium of keeping track of all the links on a page and following them one by one, and avoids accidental omissions of data.

6.4 Filtering out majority language content

In the manual sampling method, a human researcher sees every datum that is collected, and thus can discard data that is not in the correct language. However, the automated tool has no way of knowing what content lies behind the links it follows, and so will gather a great deal of extraneous data in languages that are not desired. This necessitates the fourth step, in which extraneous data is automatically removed from the growing snowball (4). In order to separate content by language, language identification software is needed. Language identification software exists for many majority languages and even some minority languages;[30] however, a digital minority language by nature will have little in the way of technological support, and so in most cases the target language will not have the ability to be identified by a machine. Nonetheless, the language identification tools written for majority languages can be used to filter out any undesired data from majority languages, leaving only content in the target language, as well as content whose language could not be determined.

6.5 Selecting target language data by hand

The final step in each wave of snowball sampling is again a manual one (5). Human experts who can identify the target language must select the content that is to be added to the corpus, after most of the pollution from majority languages has been automatically removed. Crucially,

[30] Google's Compact Language Detector, for instance, can detect around 161 languages (McCandless 2013). This is an admirable number, but still leaves many thousands of languages undetectable.

human experts are also required to evaluate every piece of video or audio content, which, barring huge advances in speech recognition technology, will long be difficult for machines to evaluate.

6.6 Continuing the snowball

The steps (2-5) can be repeated as many times as desired to increase the sample size. As the corpus is constructed simultaneously with the expansion of the snowball, the results of the sampling are useable even before the researchers decide to stop sampling.

7 Conclusion

Despite the internet's potential to aid in language revitalization efforts, little is known about the status of minority languages on the internet, and current methodologies are unable to examine this topic in a systematic way. In this thesis, I critiqued the existing literature on the internet's linguistic diversity, arguing that it privileges data on majority languages and leaves a dearth of information on minority languages. I proposed a new snowball sampling method designed to locate content in languages that would be missed by existing techniques, then applied this method in a series of case studies.

While I have only presented a brief demonstration of the snowball sampling approach, further work could explore its effectiveness through a detailed study of a single language. Extending the method for the study of sign languages would be informative not only for sign linguistics and Deaf studies, but would also provide a better understanding of how to sample video data. Long term benefits to the study of digital language use could be realized by implementing the automated system I have described, just as the creation of Praat, ELAN, and other software has enriched the field of linguistics.

8 Acknowledgments

I am indebted to K. David Harrison for his guidance throughout the creation of this thesis, and to Zhiyin Ding for her feedback on my earliest draft. I am thankful for the advice Ted Fernald and Emily Gasser provided on the final draft of this thesis. I am likewise deeply appreciative of Emma Corngold for her advice and support, as well as of my parents, Margaret and Eric Nilsson, for their feedback and critique. I would like to extend a plethora of thanks to Tamsin True-Alcalá for the feedback and solidarity she provided during the course of the thesis-writing process.

References

Anderson, Gregory D.S. & K. David Harrison. 2014. *Yokoim Talking Dictionary*. Living Tongues Institute for Endangered Languages. http://www.talkingdictionary.org/yokoim.

Babel. 1997. Web languages hit parade – June 1997. Retrieved October 11, 2006 from http://alis.isoc.org/palmares.en.html#liste_langues. Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Booth, Madeleine. About the Yokoim dictionary. In Anderson, Gregory D.S. & K. David Harrison. 2014. *Yokoim Talking Dictionary.* Living Tongues Institute for Endangered Languages. http://www.talkingdictionary.org/yokoim.

Chew, Yew Choong, Yoshiki Mikami & Robin Lee Nagano. 2011. Language identification of web pages based on improved N-gram algorithm. *International Journal of Computer Science Issues* 8(3). 47-58.

Climent, Salvador, Joaquim Moré, Antoni Oliver, Míriam Salvatierra, Imma Sànchez, Mariona Taulé & Lluïsa Vallmanya. 2003. Bilingual newsgroups in Catalonia: A challenge for

machine translation. *Journal of Computer-Mediated Communication,* 9(1). Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Council of Europe. 1992. European Charter for Regional or Minority Languages.

Crystal, David. 2000. *Language death.* UK: Cambridge University Press.

Dewey, Caitlin. 2013. How the Internet Is Killing the World's Languages. *The Washington Post*. http://www.washingtonpost.com/blogs/worldviews/wp/2013/12/04/how-the-internet-is-killing-the-worlds-languages/.

Durham, Mercedes. 2003. Language choice on a Swiss mailing list. *Journal of Computer-Mediate Communication*, 9(1). Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Facebook. Graph API. *Facebook Developers*. https://developers.facebook.com/docs/graph-api. (Accessed 2014-12-8).

Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Google. Developer's Guide - Google Web Search API (Deprecated). *Google Developers.* https://developers.google.com/web-search/docs/. (Accessed 2014-10-21).

Grenoble, Lenore A. & Lindsay J. Whaley. 2006. *Saving languages: An introduction to language revitalization.* UK: Cambridge University Press.

Guinovart, Xavier Gómez. 2003. A lingua galega en Internet [The Galician language on the internet]. In Bringas, Ana & Belén Martín (eds.), *Nacionalism e globalización: Lingua,*

cultula, e identidade. 71-88. Vigo, Spain: Universidade de Vigo. Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Heckathorn, Douglas D. 2011. Comment: Snowball versus response-driven sampling. *Sociological Methodology,* 41(1). 355-366.

Holton, Gary. 2011. The role of information technology in supporting minority and endangered languages. In Austin, Peter K. & Julia Sallabank (eds.), *The Cambridge handbook of endangered languages,* 371-399. UK: Cambridge University Press.

Karagol-Ayan, Burcu. 2007. *Resource generation from structured documents for low-density languages.* College Park, Maryland: University of Maryland. (Doctoral disseratation).

Kornai, András. 2013. Digital language death. *PLoS ONE* 8(10).

Lavoie, B. F., & O'Neill, E. T. 1999. How 'World Wide' is the Web? Trends in the internationalization of web sites. *OCLC Annual Review of Research 1999*. (Accessed 2006-10-11). http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000002655:000000059202&reqid=21527&frame=false. Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2014. *Ethonologue: Languages of the World, Seventeenth edition*. Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

Löhr, Doris, H. Ekkehard Wolff & Ari Awagana. 2009. Loanwords in Kanuri, a Saharan language. In Haspelmath, Martin, and Uri Tadmor. *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Walter de Gruyter.

Mas i Hernàndez, Jordi. 2003. La salut del català a Internet [The state of health of Catalan on the internet]. http://www.softcatala.org/articles/article26.htm. (Accessed 2006-10-11). Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Maxwell, Mike & Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. *Proceedings of COLING/ACL2006 Workshop on Frontiers in Linguistically Annotated Corpora*. 29-37.

McCandless, Michael. 2013. A new version of the Compact Language Detector. *Changing Bits.* http://blog.mikemccandless.com/2013/08/a-new-version-of-compact-language.html.

Mosely, Christopher (ed.). 2010. *Atlas of the world's languages in danger, 3rd edition.* Paris: UNESCO Publishing. Online version: http://www.unesco.org/culture/en/endangeredlanguages/atlas

No name. 2005. The language context of Pacific countries: A summary. *Directions: Journal of Educational Studies* 27(1 & 2). 134-143.

Noori, Margaret. 2011. Waasechibiiwaabikoonsing nd'anami'aami, 'Praying through a wired window': Using technology to teach Anishinaabemowin. *Studies in American Indian Literatures* 23(2). 1–23.

O'Neill, E. T., Lavoie, B. F., & Bennett, R. 2003. Trends in the evolution of the public Web: 1998–2002. D-Lib Magazine, 9 (4). (Accessed 2006-10-11).

http://www.dlib.org/dlib/april03/lavoie/04lavoie.html. Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Paolillo, John, Daniel Pimienta, Daniel Prado et al. 2005. Measuring linguistic diversity on the Internet. Paris, France: United Nations Educational, Scientific, and Cultural Organisation.

Pasch, Timothy James. 2008. Inuktitut online in Nunavik: Mixed-methods web-based strategies for preserving aboriginal and minority languages. Seattle: University of Washington. (Doctoral dissertation.)

Scannell, Kevin. 2014. Indigenous Tweets. http://indigenoustweets.com/.

Sperlich, Wolfgang B. 2005. Will cyberforums save endangered languages? A Niuean case study. *International Journal of the Sociology of Language* 172. 51–77.

Twitter, Inc. GET statuses/show/:id. *Twitter Developers.* https://dev.twitter.com/rest/reference/get/statuses/show/%3Aid. (Accessed 2014-10-21).

Wodak, Ruth & Scott Wright. 2006. The European Union in cyberspace: Multilingual democratic participation in a virtual public sphere? *Journal of Language and Politics* 5(2). 251-275. Cited in Gerrand, Peter. 2007. Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes. *Journal of Computer-Mediated Communication.* 12(4). 1298–1321.

Yahoo. 2014. BOSS Search API. *Yahoo Developer Network.* https://developer.yahoo.com/boss/search/. (Accessed 2014-10-21).