Umanistica Digitale - ISSN:2532-8816 - n.8, 2020
A. Hoenen, C. Koc, M. D. Rahn – A Manual for Web Corpus Crawling of Low Resource Languages
DOI: http://doi.org/10.6092/issn.2532-8816/9931

A Manual for Web Corpus Crawling of Low Resource Languages

[1]Armin Hoenen, [2]Cemre Koc, [3]Marc Daniel Rahn

Goethe-Universität Frankfurt, Germany
[1]hoenen@em.uni-frankfurt.de
[2]cem_koc@icloud.com
[3]marc.rahn@venturerebels.de

Abstract

Since the seminal publication of "Web as Corpus" the potential of creating corpora from the web has been realized for the creation of both online and offline corpora: noisy vs. clean, balanced vs. convenient, annotated vs. raw, small vs. big are only some antonyms that can be used to describe the range of possible corpora that can be and have been created. In our case, in the wake of the project Under Resourced Language Content Finder (URLCoFi), we describe a systematic approach to the compilation of corpora for low (or under) resource(d) languages (LRL) from the web in connection with a free eLearning course funded by studiumdigitale at Goethe University, Frankfurt. Despite the ease of retrieval of documents from the web, some characteristics of the digital medium introduce certain difficulties. For instance, if someone was to collect *all* documents on the web in a certain language, firstly, the collection could only be a snapshot since the web constantly changes content and secondly, there would be no way to ascertain completeness. In this paper, we show ways to deal with such difficulties in search scenarios for LRLs presenting experiences springing from a course about this topic.

Dalla pubblicazione di "Web as Corpus", il potenziale di creazione di corpora dal web è stato realizzato per la creazione di corpora sia online che offline: noisy vs. clean, balanced vs. convenient, commentato vs. raw, small vs. big sono solo alcuni antonimi che possono essere usati per descrivere la gamma di corpora possibili che possono essere e sono stati creati. Nel nostro caso, sulla scia del progetto Under Resourced Language Content Finder (URLCoFi), descriviamo un approccio sistematico alla compilazione di corpora per low (or under) resource(d) languages (LRL) dal web introducendo strumenti e un corso gratuito di eLearning finanziato da studiumdigitale, Goethe University, Francoforte. Nonostante la facilità di reperimento dei documenti dal web, certe caratteristiche del mezzo digitale presentano alcune difficoltà. Per esempio, se qualcuno dovesse raccogliere tutti i documenti sul web in una certa lingua, in primo luogo, la raccolta potrebbe essere solo uno snaphshot, dato che il web cambia costantemente il contenuto e, in secondo luogo, non ci sarebbe modo di accertarne la completezza. In questo articolo, mostriamo come affrontare tali difficoltà negli scenari di web

search per gli LRL, e presentiamo esperienze che nascono da un corso su questo argomento.

Pedagogical framework

The motto of the AIUCD conference in Udine 2019 was *Pedagogy, teaching, and research in the age of Digital Humanities* and we presented there an abstract describing the concept of an eLearning-based course given at Goethe University Frankfurt in summer 2019. The aim of this course was to enable students of linguistics, especially those studying smaller languages, to compile their own corpora from the web for the use in essays or theses on LRLs (including Low Resource Genres). We believe this course trains an ability which should be taught more widely in universities in the age of Digital Humanities, namely as a key competence in all areas: Sophisticated Web Search. As such, the course specializes in LRLs, which have a peculiar situation where it may be challenging to find their content among masses of content in larger languages, in closely related and similar sister languages and with restricted ranges of formats and topics. This paper gives a general guideline, with recourse to the experiences from many LRL search scenarios exercised within the course. The course itself is publicly available with eLearning lectures via https://lernbar.uni-frankfurt.de/ (Digital Humanities >> URLCoFi).

Introduction - LRL in place of a definition

LRLs are not unanimously defined ([15]) but characterized by the term. Few language resources are typically available for those languages. These resources are subject to debate and change over time. One basic component is (natural) written text in a language. For an attempt at defining which resources are basic see both [15] and [14]. For our purposes, we understand LRLs primarily as such languages for which the compilation of text corpora from the internet is difficult because of a lack of such texts or a reduced accessibility. We assume a certain correlation with the lack of other resources for those languages. A further subdivision however could be made by certain (non-exhaustive) characteristics of an LRL:

1. LRLs with *large speaker numbers*; these numbers determine a positive prospect for growth of the resources and also a high fluidity of contents.
2. LRLs with few *speakers*; typically, those languages are either endangered or threatened to become endangered; prospects for growth of resources are worse than above.
3. LRLs of *largely oral or hunter-gatherer populations*; often the typical usage or domains of internet use can be different from more literate communities, compare for instance [4], [8]; often these communities have few speakers and presumably some typical grammatical or lexical properties ([6]).
4. *historical* LRLs; obviously natural language content does not grow in most such cases.

For a search on the internet, other characteristics are likewise important. For instance, an LRL may have its own exclusive writing system (very rare but existent) like the Yi in China, in which case finding content is not rendered more difficult by masses of other documents of the same writing system. Before we look at such characteristics and their implications for searching and querying in more detail, we introduce the general LRL situation on the web and then conduct a first preparatory step for corpus compilation by using some characteristics to define which kinds of contents and languages are most likely to come up unintendedly together with or on top of a particular LRL in our focus. The scenario we thus focus on is one where we search as much content on the web for any LRL in our focus as possible.

General characteristics of the web and their implications for LRLs

Why is finding LRL content on the web hard at all? Or better, what makes it difficult? According to different statistics, more than 75% of the web are constituted by the 10 largest languages. Now, according to authoritative linguistic resources such as ethnologue.com, there are more than 7,000 languages in the world.

By numbers of native speakers, these same 10 languages as mentioned above constitute only roughly 36% of the world's population. This leads to an imbalance where the internet is not reflecting the world's population's native language composition but exhibits a clear skew towards some larger languages (of course, adding to this imbalance are the numbers of acquired second languages).

For a variety of reasons, this situation with the majority of content in so few languages can be a hindrance for page retrieval in LRLs. Consider, for instance, that nowadays webpages are often technically realized as instances of so-called Content Management Systems (CMS). Those offer an infrastructure where precomposed menu-items exist which are not always customizable or if they are, fewer (smaller) languages may be available. In consequence, much content of LRLs is forcedly mixed with menu items (and other marginal content) in one of the larger languages of the web which - to a certain extent - prohibits the otherwise often effective use of operators to exclude certain terms of larger languages (- operator on Google). Search engines ideally want to produce the most relevant results quickly. Since the largest languages are per se, by being large, statistically most relevant, this may lead to frequent irrelevant results when searching for LRL contents. There is also accidental orthographic overlap between LRLs and larger languages. Especially short words with frequent letters (and thus frequent words) tend to be good candidates for this, which is probably unfortunate since search engines do not generate results on the basis of the current web page but on the basis of a so-called index, a database where they usually (business secret) save only characteristic features of a web page with frequent words likely to be considered.

The word *bere* in Basque is the feminine possessive pronoun (her) while in Italian it is the verb to drink. *Kata* means *words* in Maori, *floors* in Turkish. The proportion of such words is typically below 10% but can contribute considerably to difficulties whenever there are larger languages using the same writing system – any one of the larger languages may overlap with a query term of an LRL one looks for. Linguistically, loanwords are more likely to come from a larger language into an LRL. In summary, there are a variety of reasons, linguistic and technical, why searching content of an LRL on the web may become difficult, *the needle in the haystack* being a suitable metaphor.

On the other hand, in the age of Big Data the benefit of retrieving so much data and compiling text corpora is larger for LRLs. Statistical training requires much data and so mining the internet may be more important for LRLs than it is for larger languages.

Consequently, prior to this paper other projects on the topic have come into existence, where the aim was to provide corpora in LRLs, albeit all suffering from some kind of barrier, mostly copyright related. Also, the corpora thence compiled are of course snapshots of their time (meanwhile the resource landscape has changed considerably for some LRLs). We would like to mention some of them since they are an obvious first place to search when looking for a specific LRL:

1. An Crúbadán: Scannell ([18]) collects lists of URLs and ngrams[1] for roughly 2200 LRLs (15.10.2019) in order to provide resources for Natural Language Processing on crubadan.org. However, due to copyright reasons, the texts themselves are not provided and some of the links may be outdated.

2. Leipzig Corpora Collection (LCC): Hosting contents in 252 languages (15.10.2019), the LCC ([10]) provides their web-crawled and processed texts via a web-interface.

3. DoBeS: This project concentrates on endangered languages and provides data for 60 of them on the web which are high-quality linguistically curated datasets. It is connected to the Language Archive of the Max Planck institute in Nijmegen, which has resources in more than 800 (15.10.2019).

Other general resources have grown to include considerable resources in many LRLs such as the Wikipedia and related projects. Holy books and missionary effort around LRLs have likewise seen large translation and digitization endeavors. The obvious limitation is that the materials are generally of only one genre and the language of many holy books often somewhat archaic. Nevertheless, homepages such as the one of the Jehovah's Witnesses possess versions and materials in many LRLs. However, having become aware of some purely technical use of their texts, their copyright is very explicit and should be read. Ideally the legal status (copyright, licensing) must be checked for each site from which texts are taken especially in non-purely-scientifically used corpora.

Apart from the Wikipedia, which is a free resource, OPUS ([19]) a parallel corpora archive offers free resources in many language pairs and provides a large list of free sources. There is another class of often free resources, namely legal governmental documents (and webpages) in LRLs. Lastly, resource lists such as the OLAC inventory provide also Links to online resources for many LRLs. Finally, there is a tool, which has been used with larger languages, but which can also be used for the automatic compilation of corpora in all languages, naturally also LRLs.

[1] *n-grams* are sequences of *n* characters or words which occur in sequence in texts, where *n* stands for a number.

This tool is called BootCat ([1]). While we do not attempt to evaluate the performance of BootCat for LRLs in general, it may suffice to say that despite being a great tool for our purposes, noise and parametric limitations such as towards the search engines used, numbers of terms and tuples etc. are reason enough to still embark on the endeavor of manually retrieving content. Users should however use BootCat and combine its results with those of manual search.

Summarizing, when building a corpus for some LRLs (or some low resource genres/ text types etc.), we believe that an integral part of looking for content must be manual content search. It can be an extension to existing resources and tools or the main key activity for corpus compilation.

Lastly, facing a medium such as the internet results are probably always at risk of being relatively quickly outdated. Some of the difficulties we face, and faced in this article, may become obsolete due to the arrival of new technologies, new devices and new regulations etc. We believe however, that the picture will likely just become more complex than changing completely, leaving some of the results still accurate, while others will be added, some outdated and some updated. We believe the internet is made of strata of webpages, some, like http://info.cern.ch/ go back to the scientific very beginning of the web, others reflect taste and technical infrastructure of the 1990ies or 2000s and yet others are recent. While our methods are based on the peculiarities of the pages until 2019, we cannot foresee further complexity added after that.

That said, one additional difficulty of any (written [or spoken]) communication is the forced linearity of language ([17]) and the missing adaptability of written language. We need to figure out a sequence of the things we want to describe and a back and forth between paragraphs is cumbersome as it usually involves large eye skips which so conveniently are largely absent from reading a print book. Our article is intended for multiple readerships, but since the authors do not have the time nor the possibility to write 10 different versions of the article each with different levels of detail, terminology and sequences, each for a different readership, the burden is upon the reader to cut his way through the forest of this article. In order to (hopefully) facilitate this, we will introduce some few tags with those types of LRLs (and readers) for which the tagged content is relevant.

1. Step - Defining Distractors

Distractors are systematically occurring instances of another language or consistent paralinguistic type of strings (such as faulty OCR, written glossolalia, machine codes, encrypted text or other artifacts) in the results when querying our target language. This must be distinguished from noise, where we understand noise to be such documents or search results which surface for a specific or a very limited number of queries only. The border between distractors and noise is fuzzy, ([18]) intending only linguistic distractors used the term "polluting" languages. Whenever intending to manually find content in an LRL we advise a first step, where the main distractors are listed in order to develop strategies to exclude their content in the results and thus increase precision of the search results carefully without affecting recall.

Related Linguistic Distractors

Linguistic distractors can again be subdivided into several types but let us first look at a characteristic of our LRL. Firstly, there are languages, which are part of a larger language family - LRL$_f$. Then, there are language isolates LRL$_i$, which have no family, a true isolate doesn't even have closely related varieties. Finally, there are mixed languages LRL$_m$ known as Pidgins and Creoles, which often exhibit some grammatical features of their substrate while adhering largely to the lexicon of the superstrate. Depending on this characteristic, a language may or may not have closely related sister languages which can be some of the most distracting distractors. Naturally LRL$_m$s are confusable with other LRL$_m$s of the same or even a closely related superstrate. The border between language and dialect is fuzzy and a matter of definition. As the famous saying attributed to Weinreich reminds us *A language is a dialect with an army and a navy.* Thus, searching for an LRL, one should be able to clearly delimit the variety against others or include different standards then delimiting the varieties from others in the same continuum. <TAGS for=”LRL$_f$, LRL$_m$”> If one has achieved that, language family trees (classificatory systems) can help us immediately identify candidate distractors. Likewise, a matter of decision is the upper node at which one considers a family a family (would we take Indo-European [for our purposes surely too large a unit], Indo-Iranian or Iranian as the root for the family of Balochi?). Generally, a closer root for the family with an uncontroversial branching point and consistent with general linguistic typology appears advisable. Online sources for classifications are sites such as Ethnologue, Glottologue, WALS, to name but a few. Here, one may search for the closest languages and then test how close they are. For them to be formidable distractors, they should have the same writing system (otherwise they can be skipped) and very few differences from our target LRL. A second factor is the size or status of the sister language, if it is a large language, one may want to put it on the distractor list even if a little less related than another closer LRL sister. There is no perfect strategy of defining a threshold for a distractor sister, but a simple heuristic may be useful: *if the percentage of exact orthographic overlap (and/or overlap in trigrams) between two as large and clean as possible word lists (the distance may be weighed by frequency) considerably exceeds the average accidental overlap with the largest unrelated languages of the web, one should include it*. One may also consider overlap in grammatical features (WALS) or phoneme inventories and graphemic systems additionally.

The exact threshold may vary per scenario, but this is not tragic, since one always has the possibility to extend the distractor list later. Again, generally one may rather include too many than too few distractors. </TAGS for=”LRL$_f$, LRL$_m$”><TAGS for=”LRL$_m$”> A side-note is that one may want to consider an LRL$_m$ attached to the root node of the superstrate language's family and likewise all other LRL$_m$ with the same superstrate. The test on the percentages of overlap may be run as described.</TAGS for=”LRL$_m$”> Historical stages (earlier distinguished varieties) of our target LRL may always be a related linguistic distractor.

Unrelated Linguistic Distractors

After having listed the closest related distractors, one may proceed to unrelated linguistic distractors. Here, the linguistic processes of loaning and borrowing[2] is the most important one for the potential of becoming a distractor. Consequently, Sprachbund and regional proximity are good estimators for candidates, but also former colonial languages. Loanwords may constitute a considerable part of a language's vocabulary. Since the direction of loaning often reflects power, it implies a larger part of words from bigger languages. The opposite direction however is equally relevant, words for local plants, meals, industrial products etc. which the larger language has absorbed. What we are looking for with non-related distractors is languages which are either the (ultimate or intermediate) source of many (loan) words of our target LRL or which have themselves borrowed a considerable number of terms from our target LRL or various mixtures of both. Loans may be mediated (for instance, directly loaned from French, but originating in English). If those terms preserve their original orthography (at least in some of the terms or to large degrees), then they can lead to serious distraction in queries. Especially if the contact is so intense, that even some (frequent) function words such as discourse markers (like *amma* in some oriental languages) have been loaned. It might also matter if one looks more for formal or informal language. All LRLs usually loan and borrow, albeit for different reasons and in different ways; LRL_ms maintain some words of their substrate for instance. Again, the question when to include a non-related LRL cannot be answered globally, but some thoughts may facilitate the decision. The larger the distractor is (especially English, which loans into many languages, likewise French, Spanish in South America, Russian in the East, Mandarin Chinese in the Far East) the more likely a distractor it will be since these languages are typically associated with plenty of web content. On the other hand, the level of perseverance of original orthography (in the original alphabet) is also very important. While Japanese has many English loanwords, rendering “glass” as グラス is hardly producing any English content (aided by the fact that in the Japanese rendering there is no more distinction between r and l, so the same transcription could also stand for the word “grass” which is however usually not used as an English loan). Here, one may thus measure how many words in a sufficiently large wordlist of the target LRL (if one has one) are English words (intersecting with a big English word list or an online corpus query API). In this case, it may also matter which words are such loans: are they all very low frequency terms, so that there is no big danger excluding target LRL content when excluding content containing them? Based on these characteristics and individual ones, one may decide to list a language as distractor language.

Paralinguistic Distractors

<TAG for=”mostly Latin alphabet based LRLs”>

[2] The common use of *loanword* and *borrowing* in English is at least partly synonymous, see Oxford Advanced Learner’s Dictionary (22.10.2019).

Finally, there are paralinguistic distractors as we learned during our course. It is hard to preview which LRL will attract which paralinguistic distractors, but based on the spread of the Latin alphabet and the fact that most programming code and mark-up code etc. is written in Latin alphabet, it is clear that the likelihood for paralinguistic distractors is clearly much higher for languages written in the Latin alphabet.[3] Now, a paralinguistic distractor should be a consistent unit if one is to later find strategies to exclude content of this type. We found some possibilities:

faulty OCR: Some websites provide text derived from OCR which was neither post-corrected, nor very accurate. In fact, there are tremendous examples, where there seem to be many more wrong characters than correct ones; and the numbers of such documents seem to be on the way to being Big Data. OCR-errors are often systematic, that is the same letter sequence in the same font tends to be misread as the same wrong sequence each time encountered. This leads to the systematic distortion of the original language of the OCR (or if the faulty parts make more than 50% of what is being displayed some language like gibberish). If one now thinks that language change likewise distorts an actual variety to form another one, it becomes intuitively clear that the distorted OCR of a sister language or even an unrelated language can accidentally (if in the process some very frequent function words and some fewer, longer content words are produced) resemble another, if we are unlucky our target LRL. So, faulty OCR is always a bullet point on the list of distractors.

glossolalia and pseudo-X: Some comedians imitate languages by producing fake sequences incorporating many characteristic sounds and maybe some words of certain languages. To give a written paraphrase, what language would you guess the following sequence to be *Das Gehortung warrende Humpelkatz rimpelt in Ratzfatz*? Does this look like German to you? Actually, as on the 10th of October 2019, Google translate would also classify the sentence as German and it probably should. Only two words and two morphemes are German: *Das* is the German article, *Ratzfatz*, better *ratz-fatz* is an onomatopoetic meaning *immediately*, the word *Humpelkatz* could be analyzed as limping cat. We hope this made clear that such imitation works also on a written basis. If we take this to be a sequence uttered or written by a comedian to humorously emphasize a rough sounding aspect of German maybe based on its affricates and some other phonotactic properties, then with some actual words produced by accident or intent such a sequence would make a formidable distractor (as also the language classification systems witness). Besides comedians, psycholinguists may use pseudo-words (with a certain permissible orthographic and phonological structure) or even phrases for experiments and some religious activities include speaking in tongues (or glossolalia) although this is seldom written down (but could be as auto generated and separated subtitles to a YouTube video

[3] Here, we would like to clarify that if one searches for transliterations, too or for multiple renderings if a language has more than one principle writing system (such as Serbian) the strategy is different. To increase clarity of our method, we assume however that the search scenario involves only one (main) writing system. If your scenario involves more, you can break it down into one scenario per writing system.

for instance). Despite being much less abundant than faulty OCR, one would want to avoid such content in a serious corpus, which is why it is to be considered a possible, yet rare distractor.

encrypted text and orthographic plays: Since antiquity ([7]) people have used codes to transmit messages which should not be intercepted. Since then an arms-race between cryptography (the process of encoding messages) and cryptoanalytics (the process of decoding) has taken place and seen the development of many different techniques. Some of these codes may produce text looking like our LRL or worse but extremely unlikely senseless text in our LRL (here, one may remember Chomskys famous sentence: *Green colorless ideas sleep furiously*, which is asemantical but in principle not agrammatical). Or one could simply write a Latin alphabet-based language in a Semitic style omitting short vowels, which then may accidentally look like another language. Also, one could invent a new orthography (for instance a simpler one for French, which then could resemble a French based creole while not being one). This distractor should also be rather rare, but there are cases where it could be a serious distractor.

machine code: programs and men produce all kinds of codes for instance in transmission or machine2machine communication. Again, accidentally, these could at least in parts look like sequences of words in the target LRL and thus become some kind of distractor.

abbreviations and acronyms: heavily abbreviated text may also lead to a completely different linguistic appearance. Since there are languages, where an abbreviation must not be marked as such (as with a dot in English), these cases could lead to another distractor. Especially in short message communication (where space is also money) innovative abbreviations have become a substandard (lol, 4u, 2b or not 2b, etc.).

Apart from these, there might be other paralinguistic distractors such as lists of names, but we were neither aware of others, nor did we witness them during the course. The extent to which such paralinguistic distractors play a role is largely dependent on the target language and on connected random factors such as the frequency of certain easily misOCRed ngrams such as ni (for m). It may therefore seem advisable to check these paralinguistic distractors one by one. In a tiny experiment, we took an abstract of a paper in Indonesian, reduced the image size and quality by a Ghostscript command, took a screenshot of it and then let tesseract OCR with English extract text from the resulting png. On the Textfile we ran fastText and found that about only 40% of lines had been identified as Indonesian, 33% as English, 15% as Malay and one line each as Finnish, Russian, Hungarian, Ukrainian and Catalan. So, in all, more than half of the lines were misclassified, while especially English and Malay, one close sister language and a contact language had larger proportions. Badly OCRed sister languages may be especially prone to become a distractor.

</TAG for=”mostly Latin alphabet based LRLs”>

With a list of distractors at hand, only two things are missing before starting the search.

Legal Issues, Copyright

Before proceeding, a word of caution is in place. Generally, determining a legal policy on web texts is both difficult and not yet internationally uniformly resolved. Usually, the law of the country of the place where the web server is located applies. For scientific use there are various rules, such as the so-called fair use doctrine which ascertains the legality of the usage of copyrighted material without permission request for educational purposes under certain conditions. A company may sue people using their texts if they suffer a loss of profit (and certainly will if this is large) - imagine a web corpus makes an annotated version of a complete Harry Potter novel available which people could use for reading instead of analysis. Likewise, small languages or religious communities may feel preyed upon when their materials get used (especially if to the end of a profit). A large project on web crawled corpora ([2]) states that:

> "The copyright issue remains a thorny one: there is no easy way of determining whether the content of a particular page is copyrighted, nor is it feasible to ask millions of potential copyright holders for usage permission. However, our crawler does respect the download policies imposed by website administrators (i.e. the robots.txt file), and the WaCky website contains information on how to request the removal of specific documents from our corpora".

Also, they argue that their corpora are processed, that is annotated which represents additional work on the texts, they are not anymore the same as the raw material. A complete download of a corpus is what we term here *legal level 1*. The LCC makes their contents available only through queries via their web-interface while a complete download is prohibited, *legal level 2*, a model which many websites with linguistic resources follow in order to avoid legal persecution (the Harry Potter novel may give examples in an analysis but the complete text is not downloadable). The An Crúbadán Website goes a step further, *legal level 3*, and does not make the texts themselves available but only word and ngram lists. Especially if one plans to publish entire text collections of texts in a target LRL from the web or to use explicit examples from such a collection (not only statistical and metadata), one should very carefully examine the legal status and options and offer the smaller language communities wherever possible a say in whether they want their texts to be used in this way or not and if profit is involved a way to participate. If using the texts for scientific purposes only one should at least make sure that this is covered for instance by the fair use doctrine. As a reference the work of the ELDAH should be considered, also as an address to turn to for specific questions.

Wordlists for query generation

Before starting to query a search engine, one needs some terms from the target language. Since we assume for our scenario that there is a need to compose a corpus, we also must assume that the researcher is not in possession of such a corpus beforehand. What follows is that (s)he must obtain a wordlist in other ways. The properties the wordlist should have are determined mostly by statistics. Such a list should feature very frequent function words, words of intermediate frequency and some rare specialized longer content words. Generally, the larger the list is the better it is and the more information on a word (its frequency or frequency class for instance) is available the better it is for monitoring and evaluation purposes. A printed grammar may be a good starting point for manually extracting terms.

Ways to search the internet and other networks

The world wide web (WWW) is an open public network of computers where some (servers) can be accessed via a curated system of addresses from any computer connected to the network for instance through a web browser. Since each of the servers can host varying numbers of pages and content and furthermore, since pages are constantly updated, removed or added (fluidity), nobody really can know how large the WWW is content-wise. There are estimates on the size of the WWW, but naturally it can't be verified. [20] estimate only the size of the portion indexed by search engines and gives roughly 6 billion pages (15.10.2019).

URL Guessing

Explained in a simplified way, to access a certain content, a user can type the address in a Browser and thereby asks the server registered under this "name" for data which it sends back. This is also the first mechanism by which to find a particular web page: direct address input. Now, if one does not know a page, one can guess names.[4] LRL communities might have names such as *language-name.country-code-top-level-domain*, for instance for German deutsch.de.[5] This may or may not work, in case it doesn't either content in another language is accessed there, a provider has reserved the name and advertises it there or the address cannot be reached. This way of obtaining LRL content is both cumbersome and has a low probability of success. This is because there are numerous possible combinations of subdomains, top-level domains and hostname letters to compose a URL (in various alphabets and ways such as Punycode) and since LRL content may only be present on some files (thus subadresses on a certain server) this again opens many possibilities.[6]

Link-Hopping, Surfing

The second possibility to search the internet is via so called hyperlinks. With any starting point in the internet, one can search and follow the links on that page and go to others. For LRLs

[4] Analogously, one may also guess IPs, which browsers also understand, but in guessing which IPs link to content in a certain LRL, only non-linguistic clues are useful for which we have no good description.

[5] This is just a quickly retrieved example, we do not want to advertize or promote contents hosted on that site.

[6] We do not treat FTP and SMTP protocols here.

this strategy is especially useful when pages are interlinked. There has been a lot of research on the topology of the internet when symbolized by a graph where pages are nodes and hyperlinks edges. One famous contribution ([3]) assumes a bow-tie model. Another famous topological property could be the small world property ([16]) where few pages function as linking hubs which connect many groups of loosely interlinked pages (which rarely link to pages outside their groups). Finally in search engine research so-called hubs and authorities ([13]) are being distinguished where hubs by and large correspond topologically to those in small-world graphs whilst authorities are pages which provide high quality content. Should the webpages of our target LRL be located in disconnected components of the WWW, meaning such groups of pages which are not interlinked with the main core of the internet, this could make them considerably harder to find. Firstly, because then there would be no way to find them through hyperlink hopping (surfing) from pages in the core. Secondly, search engines - the third way of searching the web - would be somewhat less likely to index these pages and thus they may not be retrievable through them. However, we found during the course that most pages in LRLs (of various sizes) were usually somehow connected to the core.[7] Furthermore, we realized that at least some proportion of the disconnected components could be fresh pages, which have not yet been filled with content. Thus at least some disconnected pages are irrelevant. Yet, we cannot exclude that some pages in LRLs are found in disconnected components and as a consequence are only accessible through URL guessing or informants.

Search Engines

The third and presumably most well-known way of searching content on the web is through search engines. A description of the process is problematic for various reasons, mainly because the exact functioning of search engines is their business secret and subject to change at least of parameters at least every now and then (for otherwise people would manipulate webpages in absurd ways in order to come up on top of certain customer-loaded searches).[8] In a nutshell, a search engine periodically sifts through the web (or portions deemed relevant) and generates (or updates) a so-called index, that is a database where addresses are stored along with some features. Now, because this is a secret it is not clear what these features are, but certain words (or ngrams) and their frequencies as well as the number and sources or targets of incoming and outgoing links should be involved almost certainly. On the basis of these indices, search engine queries which the user sends to the web interface of the search engine are being answered. Thus, for each query a search engine receives, an algorithm produces those result pages which according to the features of the index are most relevant to that query. Thus, one does usually not search “the internet” when using a search engine but only those portions of it known and relevant to the search engine. Generally, large search engines such as above all Google (and Bing) seem to have the largest indices. Further below, we describe how to retrieve LRL documents through search engines by composing linguistically informed queries. A side-note is that semantic web technologies (and search engines) could play an ever more important role if

[7] This refers also to pages, which we knew and found not by search engine queries or surfing.

[8] Actually, people try this through reverse-engineering. The concurrent field is Search Engine Optimization.

considering projects such as Babelnet which connect languages, some of them small, through semantic web technology. For instance, could a tag for a small language's name be connected in many meaningful ways to different types of content for it. Meta-linguistic ontologies such as OLiA can already be used to obtain features and characteristics of a potential target language and are used also for smaller languages (such as Dzongkha or Yucatec Maya or Fon) and in interlinked lexical resources.

Dark Webs

There are alternative webs which some call darkwebs. The *onion-web* is the largest of those and one of the earliest ([9]). It was intended as a place where political dissidents could voice their opinions if officially oppressed and actually parts of this web are used today for such purposes. Facebook also has a presence there. However, since in these kinds of webs, users who host or view pages are rather anonymous as secured through the technical underpinnings of the system, which at the same time make it slower than the WWW, much criminal activity is also to be found there. Since LRLs can be oppressed, content may be located in a darkweb. The way to find such content is through darkweb search engines or Hidden Wiki lists with thematically ordered entry-points, and through surfing. URL-guessing is rather impossible, since darkweb URLs are usually long codes, not registered and intentionally chosen as in the WWW (or Surface Web as darkweb surfers may call it) and seldomly meaningful. Legally and ethically, the use of texts found on a darkweb may present a greater challenge than those from the Surface Web. One has to think about their potentially threatening contents, the risk one may put their authors or oneself to using them, the unclaimed or unclear copyright situation and quotability (reproducibility) since contents appear and disappear at high rates in this segment of the web.

Social Media

One must mention Social Media such as Facebook, VKontakte and Twitter, which come with their own search engines and where accounts prompt users into areas, which the bigger search engines are not supposed to index or offer publicly. For these reasons, some content may only be accessible while logged in into an account in a Social Network and consequently looking for content in LRLs may involve Social Media Search. While as of summer 2019, Facebook closed its semantic graph search where one could for instance query all accounts of *Hindi-speakers living in Canada and working as teachers*, the current search functionalities typically still support searches using individual characteristics such as the school users attended (provided they input this information). However, the search is no more semantic but now combines query terms and is no more fundamentally different from conventional search engines.

Deep Web

A last point to mention is the so called Deep Web, which refers to content which is generated dynamically from the contents of a database only in the moment, the user onsite submits a form or performs a related interactive activity. Such content cannot be indexed by larger search engines unless they mimic the onsite query behavior which would clearly be unfeasible, given for instance all possible queries on a train company's website for the next trains at every point in time, as well as arrival and departures times at every station. It has been estimated that considerable portions of the WWW are potentially Deep Web contents and they comprise also genuine LRL content.

Search engine queries for LRLs - step by step

While guessing URLs is a possible strategy, the probabilities of success are rather low, so that it seems more advisable not to start your endeavor with this kind of search on the WWW unless for reduced ubiquitous examples. Likewise, without a good starting point (for instance one of the pages called hubs) crawling or surfing the web (hopping from page to page via hyperlinks) might be a rather difficult start let alone be cumbersome. The easiest and probably most successful way to start the web search for LRL corpora is thus the use of a search engine interface.

In this section, we describe step by step how to look for content in the target LRL by using search engines.

0. Looking for resources on known sites and BootCatting

The first place to look for content may be one of the larger sites and projects which have focused on LRLs, such as the above-mentioned ones (DoBeS, An Crúbadán, LCC, etc.). However, their content should be thoroughly checked for noise.

Secondly, using BootCat and feeding it with a limited number of terms may be a good start. The results can be URLs or even directly a corpus which BootCat draws from those. The result must then be checked manually and purified, which can in the ideal case provide a larger basis for manual queries or a second and third BootCat round. For some LRLs however, BootCat may not be able to retrieve valid content.

1. Single Term Querying

For a first approach to querying search engines we can take single query terms and compose an evaluative spreadsheet where we note for instance a) how many of the first 10 pages returned for a certain query have been in the target language and b) how many results have been returned.[9] We can then annotate different statistical or semantic or other linguistic properties of query terms. We query some very frequent function words (avoiding terms accidentally or by loaning overlapping with one of the big languages of the WWW or our distractors), some intermediately frequent terms, some content words (even if our wordlist features no explicit frequency information, we can, by universality, be quite certain that a not too unusual and

[9] Note, that the number-of-results estimate which some search engines provide is based on a quick precomputation and can deviate from the actual number of results.

short function word is very frequent, whereas a longer content word, which is not general such as *thing*, *animal*, *machine* should be infrequent). We can try words from different genres, different registers, regions etc. Afterwards, we can look at the list and try to draw inferences on the effect of the assumed characteristics on the evaluation (a and b or a [weighted] product of both). Some properties are unsurprising and can be hypothesized a priori such as function words return more results than content words or that specialized terms return less than more general ones which at times will be used as anaphora for the former. However, the current composition of content of our target LRL on the web is what has to be characterized. Guiding questions such as *does a considerable proportion of the assumed content contain pages with folklore content* could be tested by using concurrent vocabulary and seeing if result numbers or precision increase. These questions may be very individually dependent on the current time, situation and other circumstances of the LRL community at hand.

2. Multiple Term Querying

After having evaluated single terms, where some may have been found useful, the next proposed step is in combining terms. Generally, the first term can be thought of as constituting a certain amount of results and each subsequent term as eliciting a subset of the previous results, so that in principle on average the number of results decreases with the number of query terms. A function word is then a good first query term but so-called stop words should be avoided.[10]

Stop words are usually very frequent function words which big search engines simply ignore if they appear in queries. The reason is that those would simply elicit too many documents as they appear in quasi all of them (think for instance of the English article). This touches upon another issue, the settings for search. For stop words to be relevant, a search engine must have a list of them and in turn a setting for searches in the target language. Other language dependent settings may also exist and can crucially influence the results in ways again hidden in the business secret.

Often queries contain terms, which have not been indexed (that is they may be contained on the target website but not in the search engine index as a feature) or are simply not present in any target site. Now, it is the secret search engine algorithm which decides how to prioritize imperfect results. For instance, if you query 5 terms and no site is found containing all 5 in the index, should a result page which has only 3 of 5 terms but has a very good hub-score (or authority score) be prioritized over one which has 4 out of 5 terms but appears less connected and important?

This is an additional process to be thought of when querying more than one term. Apart from this, one can make another table (spreadsheet) and start combining terms systematically for their characteristics. Now, the possibilities for combination are manifold, combining a function word with a content word, a general with a specialized word etc. This allows hypotheses to become more flexible. Likewise, interpretation becomes more complex. The benefit is however,

[10] Likewise, non-stop word-function words appear unuseful towards the end of a query.

that we can get a better idea of which genres, registers etc. are especially fruitful in our scenario. What we found during the course was that if one combines content words which are too unrelated, since they do not occur together naturally, dictionaries or word lists may surface.

3. Queries with Operators

Operators are characters or character sequences which the algorithm of a search engine will treat in a predefined way when seeing them in a query. Among the most well-known and used operators may be the Google's quotation marks and minus operators, where the " operator forces the search to look for the occurrence of a sequence which may contain spaces as such ("an apple" searches pages which contain exactly this phrase, not such which contain *apple* somewhere, *an* being an ignored stop word). The - operator excludes. If we formulate a query and add - terms, we thereby exclude pages which contain the minusterm. This can be used monolingually for word sense disambiguation in queries, for instance searching for *bank* being interested in the use of the term as *riverbank*, one may formulate

bank river water -financial -money -business

in order to exclude pages where the dominant sense of the word *bank* which one can naturally not exclude from this query is the financial institution. In parallel, we can exclude other languages (big WWW languages and our distractors). Since minusterms apply only if any page at all is present in the result set, which has them, one can theoretically add a large number of them to any query. Note that search engines limit the number of maximally allowed search terms (either in tokens or as a certain bit encoding size).

4 General Remarks

Compiling a corpus from internet sources is work-intense. At least if one aims at clean corpora with very low amounts of noise or none at all. Some sections of the internet such as member only content or certain content within social media platforms or sections of the internet not indexed by a search engine are obviously not retrievable via a/that search engine. Thus, manual content search can always produce additional content for target LRLs as long as such content exists. In doing so, we found a profound knowledge of technical and linguistic underpinnings of documents on the web useful. For instance, the search for certain document types (txt, pdf, etc.) partly benefitted from different query term selection and composition strategies (in opposition to html content, we must not expect menu-items in plaintext or pdfs for instance). The personalization options of the search engine (and so does the filter bubble) modified the results partly crucially. Generally, we found search engines to be richer in content for languages spoken on the territories close to their core language than the others (English for Google, Russian for Yandex, Mandarin Chinese for Baidu). Some content was blocked from certain regions.

2 Student Search Scenarios

During the course, the following two search scenarios have been worked out and shall be given as an example here.

Nogai (Cemre Koc)

My scenario was about the search for Nogai, a Turkic language spoken by approximately 70,000 speakers in southwestern European Russia (Caucasus). The first step of my search was downloading a Nogai wordlist from crubadan.org which contained over 200 words and word bigrams in .txt-format. Then, I formulated queries by gathering and combining random words out of the downloaded txt-file. At first, I used multiple search engines but later changed only to Google which had given me 7 results in the first run (apparently Nogai newspapers and poems) whereas the others none. However, I noticed that the retrieved newspapers and poems could easily be written in other Turkic languages of their own language branch or neighboring languages like Kumyk, Karachi-Balkar or Bashkir. The main problem was to distinguish newspaper articles and poems as PDFs between Kumyk newspaper (2), Karachi-Balkar and Nogai (3), since their writing systems and lexica significantly overlap. To tackle such a low resource challenge facing the large similarity of Turkic languages in their language family branches ([11]: 81), it was necessary to collect unique linguistic characteristics such as affixes or cognates in their forms which made it easier to discern Nogai from the other languages (distractors) and to discard unwanted content (noise). The exclusion of words where the same form appeared in a distractor language (for example уьй (house)) lead to higher numbers of results in the Nogai language (first page from 1/10 to 9/10 hits). Using combinations of unique Nogai words also provided an overall higher number of hits. While мылтык (riffle) has zero hits on the first page in Google the combination with the words сары тамбыз (august) provides three hits, which include a novel of a Nogai writer (Isa Kapaev). Furthermore, it is important to mention that the combination of words of one category or topic was advisable as then the number of hits increased. Moreover, the filetype operator used for searching PDFs only influenced the results positively. In summary, I found 30 newspapers and over 30 children's books in the Nogai language in PDF format by using combinations of unique Nogai words and a filetype PDF operator which is a considerable corpus for such a small language and as such for Nogai to the best of our knowledge unprecedented.

Maori (Marc D. Rahn)

My search scenario was restricted to pdfs and concerned with Maori, a Polynesian language spoken by roughly 160.000 people in New-Zealand. To start, I looked for frequency based Maori wordlists via Google and found one[11]. From this list, I selected a smaller subset for manual work based on the following criteria:

[11] https://tereomaori.tki.org.nz/content/download/2031/11466/file/1000 frequent words of Māori- in frequency order.doc

1. The words should be high ranking, that is frequent, so as to elicit as much content as possible.

2. Since shorter words have a higher chance of accidental overlap with another language, I preferred slightly longer words (slightly less frequent). The most frequent word in Maori, "Te" ("The") was not very suitable, as it could easily appear in other languages (which it does, for example in French "you", Dative/Accusative).

3. I also preferred words with a peculiar arrangement of letters, as well as words with diacritic symbols, further reducing the chance of a random match with another language.

For instance, the sixth most frequent word "ngā", is a perfect candidate: It is very common (meaning "and"), yet not too short, could come up in any kind of text and features a peculiar combination of letters and even a diacritic symbol.

The next step in my preparation was to think of distracting languages, especially English. The Web is full of English content, especially for a country like New Zealand where it is the main language. Additionally, I wanted to avoid mixed texts and teaching resources. Therefore, I picked the words "the, and, with" to use for the exclusion of English content (blacklist terms). Despite the fact that those terms are so-called stop words, that is they are ignored by Google when searching, when excluding them they are apparently not. Lastly, as I had discovered with a few similar queries for other languages beforehand, a lot of linguistic resources do come up, both educational and scholarly. To exclude these results, I excluded the words "language" and "status" from all of my results. To then obtain my actual search results, I picked a number of terms on the Maori wordlist that fit my criteria (ngā, kua, rā, haere and tētahi) and started searching for them one by one, using the blacklist terms at the same time. I used Google for these queries, because it simply is the biggest search engine, and there is, to my knowledge, no comparable specialized search engine for the region of New Zealand or Polynesia. In addition, Google allows for a multitude of search settings: I changed the search language to English, the region to "New Zealand" and turned off personalized results. For example, one of my search queries would look like this:

ngā -the -and -with -language -status filetype:pdf

For a first impression, I chose to estimate the percentage of correct findings (PDF's in only Maori) on the first result page. A count of total results was performed manually by counting all result pages (which is more accurate than the results estimate).

The query given above returned 54 results in total (Last checked: 21.10.2019). Of the first 10 results, 9 were completely in Maori, one was a short list mainly in Maori but with English words appearing, and one was a broken link. The other queries and their results are given here below, see Table 1.

Query	**Results total**	**Results of first 10 in target language,** **Region:** ***Germany***	**Results of first 10 in target language,** **Region:** ***New Zealand***	**Results total** **(with blacklist)**	**Results of first 10 in target language,** **Region:** ***Germany***	**Results of first 10 in target language,** **Region:** ***New Zealand***
ngā	193	4	4	54	8	9
rā	241	0	1	102	8	10
Kua	182	0	0	173	0	10
haere	163	0	1	106	5	10
tētahi	118	4	9	93	10	10
Combined						
ngā tētahi	103	9	10			
rā ngā	119	5	5			
kua rā	91	6	7			
haere kua	163	7	7			
tētahi haere	112	9	9			
ngā haere	170	7	6			
rā tētahi	105	7	8			
haere rā	139	1	2			
tētahi kua	105	8	8			
ngā kua	120	8	8			

Table 1: Query documentation for Maori. All queries had an additional filetype:pdf restriction, the combined queries used blacklisted terms.

It is relatively straightforward to see that a combination of the techniques described above yields the best results. When, however, the techniques are used one by one, the selection of search terms and blacklist terms becomes more significant. Especially when stop words cannot be used (as with larger languages), when the region cannot be filtered for, or when both applies, a combination of terms can still lead to good results. In contrast, a single word query without blacklisted terms is likely to yield quite noisy results even if filtering for the right region. However, longer words with diacritics have led to cleaner results in all circumstances within this scenario.

In both scenarios, pdfs played a crucial role, but this is of course not necessarily so. Filetypes however may not only play an important role for the text type one may find (with presentation formats barely promising connected text for instance). The way in which one chooses terms may also differ depending on the filetype or type of web page, where informal language may for instance be associated with blog entries and comments more than with formal documents such as laws and constitutions. The two scenarios showed that the internet can provide texts even for small languages if one knows well how to find and distinguish them. They have provided many approaches for factors, both linguistic and paralinguistic, which play a role in manually querying content in LRLs. We also analyzed the lexical overlap between the languages involved and found that trees generated from the similarity matrices of lexical overlap in one case roughly reflected genealogy. Compare also [5] who found that language genealogy and language contact (often correlating well with geographical proximity) influence similarity (on various linguistic levels with a hint towards the lexicon). Figure 1 shows a Neighbor Joining Tree from lexical overlap in our corpus for Maori and distractors (mainly Wikipedias, filtered for most frequent English noise). It coincides almost perfectly with linguistic genealogy. For Nogai, the findings differed (with more assumed noise placing Russian in the middle).[12] While Russian is unrelated to Nogai, Bahasa Indonesia is a distant relative of Maori, English not. Despite that, both were clearly distinguishable in the maximal overlap they displayed with any of the other languages which suggests that it could be the case, that only relatively close sister languages play a crucial role as immediate distractors (which the Nogai corpus data roughly supports) whereas distant cousins and contact languages can have a similar degree of lexical similarity, lower than the sister languages.

[12] Compare also the classifier, we published on https://github.com/ArminHoenen/URLCoFi

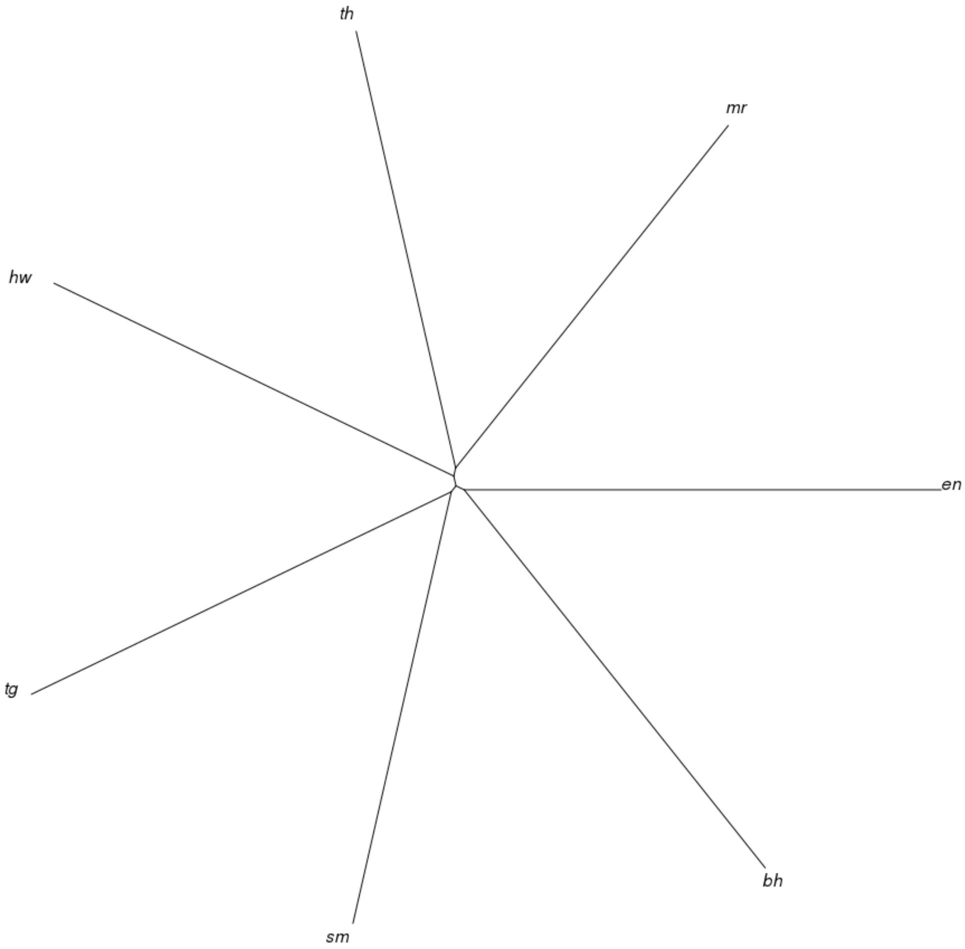

Figure 1: Neighbor Joining Tree for lexical overlap.

Experiment on Overlap

To investigate such issues further, we conduct a small experiment on the accidental overlap between an LRL and large languages of the internet. We embed this into a scenario for the decision of which distractor languages to choose. For our experiment, we take the Romance language Galician which is spoken in the north west of Spain. We obtained a wordlist from the Corga corpus[13] and extracted only the 10,000 most frequent words as we did for the largest languages on the Web (according to https://en.wikipedia.org/wiki/Languages_used_on_the_Internet: 14.10.2019, where we

[13] http://www.cirp.gal/corga/

extracted the top 10,000 terms from Wiktionary or open subtitles if the former was too small, from sources given on https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists; we did so for the top 20 languages in their first estimate plus Indonesian [so all languages in the second list are covered]; the threshold of 10,000 represents a spontaneous trade-off between frequencies and noise ratio).

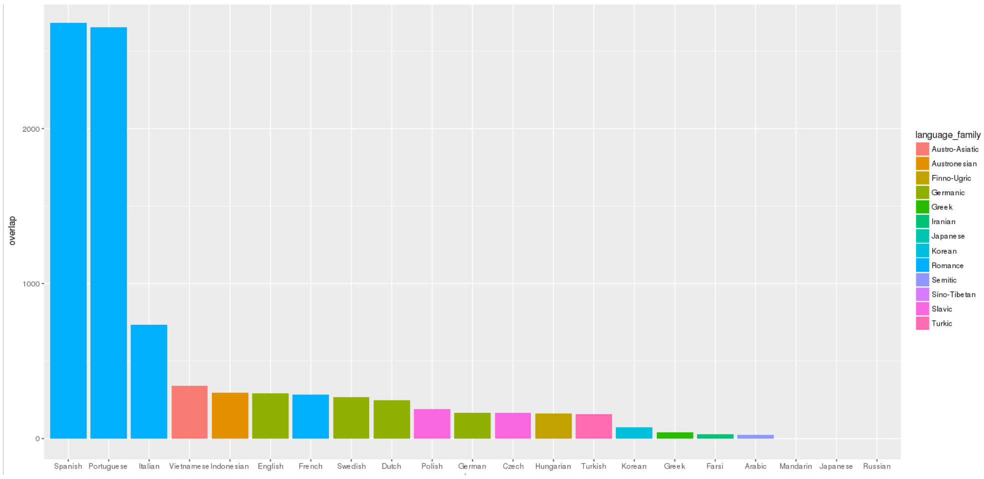

Figure 2: Number of items in word lists of important and ubiquitous languages of the WWW overlapping with the Galician word list.

Figure 2 shows the overlap per language ordered by the largest overlap. Unsurprisingly, Spanish and Portuguese have the largest overlaps. They are not only two of the largest languages on the web but also clearly the most important distractors to Galician. They have 4 times more overlap than the next language Italian, which overlaps in 733 terms. Then, French, Dutch, Swedish and English come in a group where overlap should partly be due to origin (French) or large amounts of unchanged borrowed Romance vocabulary. Vietnamese and Indonesian featured as much accidental overlap as the former group. This is mainly due to an elevated level of noise in those two wordlists. Portuguese and Spanish are clearly distinguishable by the number of overlapping terms. They should be considered distractors in this case. The distantly related Indo-European cousins which are also contact languages fell thus into one group. Compared to the Maori scenario however, the role of Italian as an intermediary here points to more variety in scenarios which might make it necessary to conduct such an overlap study in each target language case, yet with the caveat, that there might be no target language data in the first place - so it seems rather advisable to include more than fewer distractors, also from a statistical point of view.

Attempting to assess the question, if generally unrelated languages with smaller phoneme inventories and simpler syllable structures feature much more overlap, we considered all languages in the WALS which had a simple syllable structure annotated and a small consonant or vowel inventory [or both] (if requiring all 3, the number of languages decreased to two, Pirahã and Tacana – for both of which we could not locate word lists longer than roughly 1500 tokens). From these 38 languages, we found 4 to have a Wikipedia from which we extracted wordlists of the most frequent 10,000 tokens (using WikiExtractor.py for text extraction from the Wikimedia dumps from 01.10.2019): Guaraní, Hawaiian, Maori and Yoruba. Hawaiian and Maori are relatively closely related. We then intersected the lists and excluded English, Portuguese, Spanish and French words (50k lists from open subtitles[14]) and punctuation. We then added Basque which according to WALS has average inventories and a complex syllable structure to see if there was less overlap for the former languages with Basque than with each other (apart from related Hawaiian and Maori). We found hints for this although not in larger magnitudes, but a much larger investigation has to be conducted to confirm or disprove such claims and for the quantification of such effects.

Conclusion

We have presented a guideline to searches for content in LRLs on the web which sprang from the experiences made and resources gathered during a course in 2019, the concept of which we had presented as an abstract at the AIUCD 2019. The guideline included a wide variety of suggestions for dealing with manual searches for LRL content in the fluid medium of the internet and considered ways to search, tools, web and language statistics, well-known linguistic and metalinguistic sources, legal caveats and much more.

References

[1] Baroni, M., and S. Bernardini. 2004. "BootCaT: Bootstrapping Corpora and Terms from the Web." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, edited by M. T. Lino, M. F. Xavier, F. Ferreira, R. Costa, R. Silva, 1313–1316. Paris: European Language Resources Association (ELRA).

[2] Baroni, M., S. Bernardini, A. Ferraresi and E. Zanchetta. 2009. "The WaCky wide web: a collection of very large linguistically processed web-crawled corpora." *Language Resources and Evaluation* 43, no. 3: 209–226.

[3] Broder, A., et alii. 2000. "Graph structure in the web," *Computer Networks* 33, no. 1–6: 309–320.

[4] Cocq, C., and K. P. Sullivan. 2019. *Perspectives on Indigenous writing and literacies*. Leiden: Brill.

[14] https://invokeit.wordpress.com/frequency-word-lists/; we also intersected all 59 wordlists from 2016 from this project and found through generation of a Neighbour Joining Tree from a distance matrix that the data by and large reflected genealogical relationships if the alphabets were the same and found corroboration that Vietnamese and Indonesian had higher levels of noise.

[5] Cysouw, M. 2013. “Disentangling geography from genealogy.” In *Space in Language and Linguistics: Geographical, Interactional, and Cognitive Perspectives*, edited by P. Auer, M. Hilpert, A. Stukenbrock, B. Szmrecsanyi, 21–37. Berlin: De Gruyter.

[6] Cysouw, M., and B. Comrie. 2013. “Some observations on typological features of hunter-gatherer languages.” In *Language Typology and Historical Contingency*, edited by B. Bickel, L. A. Grenoble, D. A. Peterson and A. Timberdale, 383–394. Amsterdam: John Benjamins.

[7] Dooley, J. F. 2018. *History of Cryptography and Cryptanalysis*. Berlin: Springer.

[8] Evans, N. 2010. *Dying words: Endangered languages and what they have to tell us*. Hoboken: John Wiley & Sons.

[9] Gehl, R. W. 2018. *Weaving the Dark Web: Legitimacy on Freenet, Tor, and I2P.* Boston: MIT Press.

[10] Goldhahn, D., T. Eckart and U. Quasthoff. 2012. “Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages.” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, edited by N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, 759–765. Paris: European Language Resources Association (ELRA).

[11] Johanson, L., and É. Á. C. Johanson. 2015. *The Turkic Languages*. London: Routledge.

[12] Kilgarriff, A., and G. Grefenstette. 2001. “Web as corpus.” In *Proceedings of Corpus Linguistics 2001*, edited by P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja, 342–344. Lancaster: UCREL.

[13] Kleinberg, J. M. 1998. “Authoritative sources in a hyperlinked environment.” In *Proceedings of the ACM-SIAM symposium on discrete algorithms,* 668-677. Philadelphia: Society for Industrial and Applied Mathematics.

[14] Kornai, A. 2013. “Digital Language Death.” *PLOS ONE* 8, no. 10: 1–11. DOI: 10.1371/journal.pone.0077056

[15] Krauwer, S. 2003. “The basic language resource kit (BLARK) as the first milestone for the language resources roadmap.” In *SPECOM'2003. Proceedings of the International workshop (Moscow, Russia, 27-29 October 2003)*, edited by R. Potapova, 8–15. Moscow: URSS Publishing Group.

[16] Milgram, S. 1967. “The small world problem,” *Psychology Today* 2, no. 1: 60–67.

[17] Ong, W. J. 2012. *Orality and Literacy.* London: Routledge.

[18] Scannell, K. P. 2007. “The Crúbadán Project: Corpus building for under-resourced languages.” In *Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop*, edited by F. Cédrick, H. Naets, A. Kilgariff, G.-M. De Schryver, 5–

15. Louvain: Presses Universitaires de Louvain.

[19] Tiedemann, J. 2012. “Parallel Data, Tools and Interfaces in OPUS.,” In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, edited by N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk and S. Piperidis, 2214–2218. Paris: European Language Resources Association (ELRA).

[20] Van den Bosch, A., T. Bogers, and M. De Kunder. 2016. “Estimating search engine index size variability: a 9-year longitudinal study.” *Scientometrics* 107, no. 2: 839–856.

Last access URLs: 22 October 2019.