

Contextualising Levels of Language Resourcedness affecting Digital Processing of Text

C. Maria Keet^a and Langa Khumalo^b

^a Department of Computer Science, University of Cape Town, South Africa.

Email: mkeet@cs.uct.ac.za.

^b South African Centre for Digital Language Resources, North-West University, South Africa.

Email: langa.khumalo@nwu.ac.za.

Abstract

Application domains such as digital humanities and tool like chatbots involve some form of processing natural language, from digitising hardcopies to speech generation. The language of the content is typically characterised as either a low resource language (LRL) or high resource language (HRL), also known as resource-scarce and well-resourced languages, respectively. African languages have been characterized as resource-scarce languages (Bosch et al. 2007; Pretorius & Bosch 2003; Keet & Khumalo 2014) and English is by far the most well-resourced language. Varied language resources are used to develop software systems for these languages to accomplish a wide range of tasks. In this paper we argue that the dichotomous typology LRL and HRL for all languages is problematic. Through a clear understanding of language resources situated in a society, a matrix is developed that characterizes languages as Very LRL, LRL, RL, HRL and Very HRL. The characterization is based on the typology of contextual features for each category, rather than counting tools, and motivation is provided for each feature and each characterization. The contextualisation of resourcedness, with a focus on African languages in this paper, and an increased understanding of where on the scale the language used in a project is, may assist in, among others, better planning of research and implementation projects. We thus argue in this paper that the characterization of language resources within a given scale in a project is an indispensable component particularly in the context of low-resourced languages.

Keywords: LANGUAGE RESOURCES, LOW-RESOURCED LANGUAGES, CORPORA, HUMAN LANGUAGE TECHNOLOGIES, NATURAL LANGUAGE PROCESSING, LANGUAGE AUDITS, DIGITIZING RESOURCES.

1. Introduction

Digital Humanities (DH) projects span a wide range of activities and, notably, also languages. An example is the digitisation project of the notebooks in the |xam and !kun languages (Southern Africa) as part of the Llarec project¹. Similar projects take place on other continents, such as the development of digital tools for the Dongba pictogram language (southern China) (Xu, 2023), and Ottoman Turkish (Kirmizialtin and Wrisley, 2022). Neither language is in the same league as English when it asks for OCR tools for digitization, for instance. In fact, tooling for the languages were minimal to non-existent, hampering faster progress in DH research because the infrastructure hurdles have to be overcome as well. There has been a litany of terms used to describe human languages according to the data resources that they have available for the development of their human language technologies. In the context of African languages, terms like under-resourced, resource-scarce, less-resourced, low-resourced, or low resource

languages (LRL) have been used to characterize the lack of data resources, which preclude or slow down the development of computer-based and data-driven technologies in these languages. Furthermore, there is an assumption that all African languages are in the same category as LRL, while conversely a similar assumption obtains, that English and other European languages are high resourced languages (HRL). In this paper we provide a deeper understanding of this complex phenomenon.

The reference to African languages as LRL (Bosch et al. 2007; Pretorius & Bosch 2003; Keet & Khumalo 2014), while to a certain degree accurate, does not necessarily apply the same way to all the African languages. To appreciate this complexity, we need to understand what language resources entails. From the computational side, the natural language processing (NLP) literature refers to LRL as having limited online corpora and computational grammars. Alternatively, it is formulated in terms of missing something, e.g., “[...] lacking large monolingual or parallel corpora and/or manually crafted linguistic resources sufficient for building statistical NLP applications”ⁱⁱ, as if statistical NLP is the benchmark, or “LRLs can be understood as less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denomination”ⁱⁱⁱ (Tsvetkov, 2017). Compared to what? Less than which other language? For instance, Dutch is a well-resourced, but ‘less’ computerized, language than English and it has much fewer speakers. Clearly, following these measures, then every language other than English falls in the ‘less than’ category. These indications are coarse-grained and do not assist in categorizing which language is where on the spectrum. To that end, several attempts have been made by counting a selection of resources and devising a metric for classification, such as counting labelled and unlabelled corpora (Joshi et al., 2020) and availability of data and tools for a list of NLP tasks (Berment, 2004; Krauwer, 2003). But is the combination of the number of resources and their availability the only dimension of a low resourced language? We argue that it is not.

This is sharply illustrated in various half-measured attempts to bridge the gap between the LRL and English. In international collaborations for instance, the attitude is one of “just get the translators” to create the parallel corpus between language x and English, or “just get a few students to scan the texts” to create the monolingual corpus and digital library, or “quickly list how adjectives work in your language”, and that stance at times in print as well (Hämäläinen, 2021). What transpires on the ground is that, even with money available, there may be insufficient translators available, or they are not specialised in the domain one needs the translations, or they are not translators by training. And when there are enough people for scanning text, the ‘best’ OCR software is ‘best’ only for the language it is optimised for^{iv},

which is not the language being scanned and so the scanned text is riddled with errors. One can start over manually typing it up or find a computer scientist to develop language-optimised OCR. This meanwhile has been done for the 11 South African languages (Hocking & Puttkammer, 2016) – after the IsiZulu National Corpus development was already delayed due to scanning errors (Khumalo, 2021). Both are setbacks and cause further delays in trying to close the gap. For the grammar question, it may require a full literature search and compiling content from different, often outdated, sources that may not have been tested on a sizable number of words, rather than the well-resourced languages’ assumption that there is an up-to-date grammar book that only needs to be fetched from the bookshelf and digitised. From the foregoing, there are more facets to LRLs than simply not having a lot of digital language material.^v Put differently: the so-called ‘lived reality’ of working with LRLs is not one of simply missing digitized corpora and grammars.

The imperative exists, therefore, that we should have a comprehensive understanding of the language resources. This enables researchers to appreciate what the presence or absence of a particular resource entails in the characterization of a particular language. Acknowledged, the characterization of HRLs takes a similar sweeping trend. The assumption that all European languages are as highly resourced as the English language is flawed, especially in the absence of a clear matrix that assist in such a characterization.

In this paper we focus on three key issues regarding LRs: what are really the distinguishing characteristics of LRLs (and, by extension, ‘non-LRLs’), what are the characteristics of levels of resourced-ness, and which language fits where and why? To answer these questions, we first devised a list of 11 key dimensions for languages and then assessed which of those applies to which level or resourced-ness. This characterisation provides a more comprehensive understanding of the multi-faceted concept of resourced-ness of a language beyond the counting of speakers, tools, and corpora, with a focus on teasing out the details of ‘low’ resourcedness. We then took the classification and tried to allocate all 11 official South African languages, as well as several other African languages and languages spoken elsewhere in the world. This illustrates the classification is operationalizable and may further assist, e.g., language policy planning.

Section 2 discusses in more detail the typical descriptions of LRLs used in the literature. Section 3.1 of this paper discusses the matrix used for characterizing language resources. Section 3.2 analysis the features assigned to the various categories in the characterization of languages as Very LRL, LRL, RL and HRL. It will be shown that while most (European)

languages are classified as HRL, English is leading as a Very HRL, a characterization that hitherto did not exist. We discuss in Section 4 and conclude in Section 5.

2. Related work

In the fields of natural language processing, computational linguistics, and human language technologies, i.e., from the computing side, the aforementioned ‘lacking ...’ view on LRLs is often repeated in various wordings across publications, where research papers adopt such descriptions from survey papers and possibly from online media as well^{vi}. Besacier et al.’s survey (2014) describes LRLs as “a language with some of (if not all) the following aspects: lack of a unique writing system or stable orthography, limited presence on the web, lack of linguistic expertise, lack of electronic resources for speech and language processing, such as monolingual corpora, bilingual electronic dictionaries, transcribed speech data, pronunciation dictionaries, vocabulary lists, etc.”, which is roughly repeated in Ranathunga et al. (2021) and from both surveys to hundreds of research papers. Besacier et al. based their description on the Basic Language Resource Kit (BLARK)^{vii} “as the minimal set of language resources that is necessary to do any precompetitive research and education at all”, described in (Krauwer, 2003), and a similar approach based on counting resources by Berment (2004). Berment distinguishes between low, medium, and well-resourced, and provides a list of types of resources – dictionaries and a range of software applications for multiple NLP task – to assist counting them for a given language, which is then used to compute a value for categorisation (pp20-22). Krauwer provides a longer list of language- and speech technology tasks and modules with data and application availability and their importance for the NLP task, such as parsing and referent resolution (written language) and prosody and pronunciation lexicon (speech) on corpora and information access applications. The availability status of the resources collected is then estimated on a 10-point scale and added up so that eventually the resourced-ness of languages may be compared.

Joshi et al. (2020) took a narrower scope on resourced-ness assessment than Berment and Krauwer, by counting the resources available on the ELRA map for labelled dataset, the LDC catalog for corpora, and Wikipedia pages as an indication of unlabelled data in order to devise six levels of resourced-ness which is, to the best of our knowledge, the most fine-grained list on resourced-ness of languages. They gave them the descriptive names of the left-behinds (n=2191; a.o., Dahalo), scraping-bys (n=222; a.o., Cherokee, Fijian), hopefuls (n=19; a.o.,

isiZulu, Irish), rising stars (n=28; a.o., Indonesian, Afrikaans), the underdog (n=18; a.o., Russian, Dutch), and the winners (n=7; a.o., English, Japanese). Given an estimated 7000 spoken languages, there would be an implicit 7th category for the remaining 5000, the ‘not even considered’, behind the left-behinds. Assessing their taxonomy^{viii}, that latter group includes, among many, Mpiemo (for which a POS tagger does exist (Hammerstrom et al., 2008)), Pokomo, and Sanga of the Niger-Congo B (‘Bantu’) family of languages. The latter also illustrates a shortcoming of the ‘counting resources’ approach to defining and categorizing LRLs: there are always resources that are overlooked, which is due to a range of reasons interfering with the resource collection efforts and how searches for resources are conducted. For instance, multiple languages have several names or spelling variants and their use may be buried in a data collection paragraph in the paper when only titles and abstracts are scanned (Keet, 2022). Detecting mentions, like Joshi et al. did, will have missed a paper on Mboshi/Embosi, for the language is listed in their taxonomy only as Mbosi without spelling variants.^{ix}

The approach of counting a selection of resources of LRLs, and, hence, the implicit lack of things, is a recurring theme for categorising LRLs, although at times LRLs are informally described slightly more broadly. Hedderich et al. (2021) broadens LRL as an “umbrella term” that “covers a spectrum of scenarios with varying resource conditions”, which include “threatened languages”, languages that are “widely spoken but seldom addressed by NLP research”, and “non-standard domains and [NLP] tasks” (Hedderich et al., 2021). Yet, then they narrow it down when teasing out aspects of resourced-ness with their three dimensions: of availability of labelled data relevant for the task at hand, of unlabelled language or domain-specific text, and of “auxiliary data” such as a relevant knowledge base in the target language to enhance the NLP or a machine translation system to generate training data. A key observation is that indeed there are dependencies across NLP tasks and that absence of such auxiliaries hampers progress of the intended task. Occasionally there are hints at more features of LRLs beyond data and software tools, as in “less studied, resource scarce, less computerized, less privileged, less commonly taught, or low density, among other denominations” (Magueresse et al. (2020), based on Cieri et al. (2016)), but it is then still narrowed down for their works to refer to “languages for which statistical methods cannot be directly applied because of data scarcity.” (Magueresse et al., 2020) or “those that have fewer technologies and especially data sets relative to some measure of their international importance.” (Cieri et al., 2016).

Within the South African context, the characterization that emerges from the two Human Language Technology (HLT) Audits in 2011 and 2018 (Moors et al., 2018; Sharma Grover et al., 2011) is the one that refers to African languages (which are LRL) as generally “under-resourced” or “lesser resourced”. This is understood to mean that these languages have insufficient resources. The resource types covered in these audits range from the text/speech corpora to speech-to-speech translation systems. The terminology “low resourced” and “under-resourced” or “lesser resourced” may seem thus synonymous and therefore used interchangeable, however we prefer the former as it is scalable, while the latter is just a characterization of “the absence of or lesser than”.

In sum, the literature provides neither a clear definition of what LRLs really are nor what exactly their characteristics are beyond the ‘lack of’ and ‘less’ resources for computation.

3. A more comprehensive view on low resourced languages

While it is indeed the case that there are ‘few’ resources for LRLs, the few resources are not all alike and the LRLs are embedded in a broader context that makes the efforts of getting out of the ‘low resourced’ status also not all alike. Further, it requires a characterisation of its own as compared to reducing it to the negation of high-resourced. To this end, we first compiled a list of dimensions of what low resourced and what high or well-resourced entails, which is described and illustrated in Section 2.1. Second, we allocate applicable dimensions to a 5-point scale from very low to very high resourced and to what extent they apply and operationalise that for a section of languages, which is described in Section 3.2.

3.1 Dimensions of resourced-ness

The list of dimensions is summarized in Table 1 and elaborated on in the remainder of this section in that order. The first eight dimensions are described in no order and focus on multiple key aspects and consequences related to LRLs. This then led to the question: ‘what do high, and very high resourced languages have that we would like to have for LRLs too?’. Whether that is framed by an individual as envy or a wish list is neither here nor there; the aim here is to bring those benefits or features to the fore. They are summarized in features 9-11. We elaborate on each one in turn.

Limited grammar refers to the situation where a limited amount of the grammar is known, the limited documentation available may be outdated, and the few rules known may not have been evaluated on whether they hold generally or always or have exceptions. It also includes the case where there is no grammar reference book for the language in question that has the grammar all in one place, or at least of what it is known, in one place. What this practically means is that any grammar feature requires careful investigation through consulting various scientific literature, textbooks, or colleagues to distill what the rule(s) may be, and a certain amount of testing with sample data to determine whether the candidate rule(s) hold. If so, then it is still to be computerized, which may add a new hurdle: a popular academic framework such as Universal Dependencies (de Marneffe et al, 2021) may not be readily usable or useful for the LRL in question because it relies on the lexicalist paradigm whereas the language may need annotations at the sub-word level and therewith new investigation into extensions and software support for that, as, e.g., Park et al. (2021) did for St. Lawrence Yupik.

No/very limited corpora (<100K tokens) concerns both monolingual and parallel corpora with another language. The most challenging position to start from is when there is very little written text at all. Since the emergence of social media, however, this will be less likely the case, but such data requires more effort to create a corpus from, such as code switching in text messages and privacy issues. When there are texts available in digital format, it still has to be put together into a relevant corpus. While this is a relatively easy task, resources are needed to carry it out and sustain it: people to do it, funding to pay them, a plan for longevity, and open access to the corpus. Common first steps to putting a corpus together are the freely available bible translations, no matter that they are old translations and in a very specific genre that would typically not be the same as the target genre—when there are no to very limited corpora, one does not worry about genres. An experiment by Ndaba et al. (2016) showed that even training on a tiny corpus of about 20000 tokens from current news items already outperformed a language model trained on the bible in a tool for spellchecking of modern isiZulu. Then there are the OCR issues in digitization, as mentioned in the introduction.

Little basic HLTs. With basic HLTs, we mean entry-level digital support features, such as spellcheckers, an autocomplete function for tools like WhatsApp, and an online bi-directional dictionary between words in the target language and those of a more widely-spoken one.

Few people driving the advancement of the language. This is sub-divided into few researchers in languages and linguistics as well as in computer science and engineering that have the language as a research focus, few companies that have people working on the language

technologies or providing language services, and insufficient language teachers at especially primary and secondary schools. The shortage of researchers hinders progress not only in language understanding but also in prospects to devise novel algorithms to develop HLTs that work for the language's peculiarities, since the international tools typically either have to be adapted or started from scratch (e.g., Steinberger et al. (2011) and De Pauw and Wachaga (2007), among many). Few people in industry, on the other hand, is an indication that the language may not be economically profitable even though many people may speak the language. The shortage of language teachers in schools brings its own set of challenges, where learners in higher grades may not have their home language as language of teaching and learning, and, more importantly, without the language teachers, they will not learn their mother tongue (home language) well. This, in turn, has a knock-on effect on students entering tertiary education, who are then relatively behind in language knowledge, which then means either the degree programme will contain comparatively less advanced content compared to, say, the English programme at a university in the UK, USA, Australia, or New Zealand, and thus results in a comparatively less-skilled language graduate, or it is comparatively harder due to the need to catch up, which would leave fewer student to choose such a degree and more to drop out. Either case will again reduce the pool of candidates for academic posts to conduct research into the language or to teach a new generation of language teachers; i.e., it is a vicious circle at best and a downward spiral at worst. Additional effects of limited language education are that it makes it harder to recruit competent participants in evaluations for one's research and sub-competent evaluators add noise to the development of algorithms to develop the tools that may then need another round of editing.

Little/no funding applies to all the dimensions, such as funding to do research into it, funding for offering it as a subject in schools, for teacher training, and for tool development. It entails that most advances are made by passionate volunteers, who have different incentive mechanisms and may have different preferences on what HLT to invest time in than funders may have. It also affects the longevity of the outputs produced in a project: when the limited funding comes to an end, the tool will likely not be maintained or the domain name not renewed and thus the effort will slide into disuse, for someone else to re-do the same or a similar effort later for lack of availability of the preceding outputs. This will keep a language going in circles with basic HLTs, rather than advancing the state of the art. An example can be seen with isiZulu spellcheckers: they were developed in the 2000s (De Schrijver and Prinsloo, 2004) but the tool and code were never released, then from scratch again as plugin to OpenOffice^x but only for older versions (before 4.x) and therewith became defunct in the early 2010s, a stand-alone jar

file in 2016 based on new algorithms of (Ndaba et al., 2016) to avoid text editor version rot and application lock-in, and again in 2022 by CTeXT as a plugin to the Microsoft Office Suite on the PC only^{xi}.

Limited education. This relates to the previous discussion on few teachers in praxis, but here we refer to specifically the education system itself and the programmes that are available. Specifically, on whether the language is used as language of learning and teaching, on whether one can take it as a subject at school as an ‘additional language’ subject, and whether it is even possible to study it as a degree at a university. For instance, the University of Cape Town only offers isiXhosa as a major, but there is no full BA in languages, literature, or linguistics^{xii}, and the information is provided only in English, which is a similar situation at other South African universities for other languages^{xiii}. Compare this to, e.g., studying the minority language Irish at Trinity College Dublin, Ireland, where the information page is already in Irish^{xiv} and there are two 4-year degree programmes. Related to degree offerings is, perhaps, a first step on attitude to languages at a university, where there may be “language centres” to learn multiple languages at conversational and certificate levels as extracurricular activities; e.g., the small university Free University of Bozen-Bolzano’s language centre offers courses in nine languages^{xv} and Wageningen University in the Netherlands offers language courses in 15 languages^{xvi}. The UCT language centre, in contrast, only offers several versions of English^{xvii}; the University of Stellenbosch offers English, Afrikaans, and isiXhosa^{xviii}. For LRLs, even such options thus may be limited to non-existent.

Limited/no government policy framework. While arguably there is a difference between having a policy framework for a language and actually implementing it, the state of affairs for a language is worse if there is no language policy at all. In the South Africa context, the Constitution (1996) provides for the development of the 11 official languages in order to achieve “parity of esteem” between them. There have been various language policies that have been promulgated to support this constitutional imperative. However, these have been criticized as “[...] pious articles of faith, without any force nor any effect”^{xix} because of the absence of clear implementation strategies that can translate the spirit and letter of the language policy into a practical action plan that is implemented by schools and tertiary institutions. The jury is still out on how the Department of Higher Education and Training (DHET) will provide for the implementation of its Revised Language Policy Framework for Public Higher Education Institutions, which took effect in January 2022.

May not have a monolingual dictionary. We distinguish between bi-lingual dictionaries to have English/French/Spanish/Portuguese/etc. translations to the LRL vocabulary and the

recursiveness of a monolingual dictionary. The latter is seen as a measure of a certain level of resource development, status, and use, in that the vocabulary indeed can be explained in the language itself rather than having to resort to approximations to a foreign word and the LRL is thus rich enough in expressiveness to do so, i.e., to be able to describe one's language in one's own terms. This, perhaps, also concerns a legacy of colonialism and acting against it: indigenous languages were assumed to be not full-fledged languages because a people speaking it might have lived 'primitively' in the eyes of the coloniser and a monolingual dictionary then becomes a statement of belonging to humanity on par with other peoples. It also requires more effort to develop a monolingual dictionary compared to word lists, it entails that there are at least some people driving the language forward, and that there is a notion of standardization of the language and therewith institutionalisation. Note that not having a monolingual dictionary does not imply that the language is not expressive enough.

Choice among grammars is the first dimension/feature worded from the perspective of better resourced languages. This dimension of being able to choose a computational grammar for one's need or preference includes both choice between grammar paradigms (e.g., dependency vs phrase-structure grammars) and within-paradigm amongst specific frameworks, like UD (de Marneffe et al., 2021) versus SUD (Gerdes et al., 2018) dependency grammars.

Large corpora, with genres. From an LRL perspective, a corpus of 1 million tokens of anything that can be found is impressive; for better researched languages, this would be more like at least 500 million to 1 billion tokens. In addition, they would be large corpora already separated by genre, such as news articles or novels, and there may be geographic distinctions for investigating whether that makes a difference, such as American English vs British English or French from France and from Québec in Canada, German in Germany vs German in South Tyrol in Italy. Such fine-grainedness enhances not only research outputs in corpus linguistics, but also HLTs built from them.

Lots of tools and choice among tools, on multiple devices and OSs. Unlike the serial spellcheckers for isiZulu illustrated above, one rather has an abundance of choice of tools for the same operating system as well as across software, such as not only for MS Word or only copy-paste into a Web app and copying the corrections back in, but they have spelling checking embedded in the tools upon installation, regardless the device, and regardless the operating systems. For instance, that it works on both Android and iOS smartphones, on browsers when writing an email in Gmail or a blog post, in the Thunderbird Desktop email application on Windows, MacOSX, and Linux or even all platforms, and any popular text editor. In other words: there are multiple tools that offer the same functionality across that spectrum. This also

includes choice of tools for NLP tasks, such as the Stanford POS tagger and the POS tagging module that comes with NLTK, among many.

Table 1. Summary of the dimensions and components thereof, where applicable.

Number	Topic	Component-topic
1	Limited grammar	
2	No/very limited corpora	
3	Little basic HLTs	
4	Few people	4a. researchers 4b. industry 4c. teachers
5	Little/no funding	
6	Limited education	6a. as LoLT 6b. as additional language at school 6c. as degree offerings at university
7	Limited/no government policy framework	
8	May not have a monolingual dictionary	
9	Choice among grammars	
10	Large corpora, with genres	
11	Lots of tools and choice among tools, on multiple devices and OSs	

The criteria or dimensions characterized above form a matrix that we will use in section 3.2 in analysing the complexity of understanding LRs. We will show in section 3.2 that the characterization of languages from Very LRL to Very HRLs is based on well worked-out criteria or dimensions, with examples of such languages provided. The characterization Very LRL, LRL, RL, High RL and Very HRL moves from a subjective to a more scientific/structured/motived criteria.

3.2 Analysis

The analysis uses feature-based criteria. The features are assigned to each of the five categories Very LRL, LRL, RL, High RL and Very HRL. The 11 official languages of South Africa are assigned to categories in 3.2.2 together with other African and global languages.

3.2.1 Matching dimensions to the level of resourced-ness of a language

A basic categorization, or levels, of resourced-ness easily adjusts toward a familiar 5-point Likert scale version. For resourced-ness of languages, that are first the well-known “very LRL” and “LRL”. What comes after is also not one homogeneous group, however, and with an averagely resourced language (“RL”) in the middle, the matching other side it is highly or high resourced language (“HRL”) and the extreme on that end, a very high resourced language (“very HRL”). We assign the features to the five levels as follows:

- Very LRL: 1, 2, no 3, 4, no 5, 6b, 6c, 7, 8
- LRL: 1, 2, basic 3, 4a, 4b, sparse 4c, 5, 6b, 6c, 8
- RL: 3, 5, may have one of {9, 10, 11}
- HRL: some 9, some 10, some 11
- Very HRL: 9, 10, 11

This is also visualised in Table 2. What is especially clear from the visualisation, is that the difference between LRL and RL seems to be the largest. ‘Crawling out of the pit’ of relative disadvantage to ‘getting by fine’ may indeed well be the hardest to do, since once the system is rolling, it makes it easier to build upon it and move forward faster, but getting it rolling takes more effort.

What also shows is that “Limited LOLT” (6a) turned out not to be a problem. The (very) LRLs are used in education, and especially in primary schools, and of course this happens as well in the resourced languages. Reflecting on why it was added, was because the LRL is often perceived as lower status and replaced by a colonial language as LOLT especially in higher grades and better schools, but not for all LRLs. Hence, it could not be applied to LRL or very LRL as a category, but only to a language specifically.

Table 2. Visualisation of the dimensions by level of resourced-ness. X: yes/applies; v: applies to a considerable extent i: applies to some extent; blank cell: does not apply.

	1	2	3	4			5	6			7	8	9	10	11
				a	b	c		a	b	c					
Very LRL	x	x	x	x	x	x	x		x	x	x	x			
LRL	x	x	v	x	x	v	x		x	x		x			
RL			x				x						i	i	i
HRL													v	v	v
Very HRL													x	x	x

3.2.2. Categorising languages

With the aim toward operationalizing the dimensions and levels of resourced-ness and illustrate the categorisations and ‘lived reality’ better, we now turn to examining which language can be categorized where and why. We do this specifically for the 11 official languages of South Africa, several other more or less well-known languages in Africa, and a selection of other languages in the world for comparison and to provide ample examples in each category. There are many ways how such a categorisation can be done, which each have their own shortcomings. For instance, one may try to collect all the resources and tools ever developed for the language or only those that are functional (be it assessed or claimed on a paper or open access), or collect papers about resources from a wide range of publication venues. Such resource collections and audits have been carried out, including, the HLT audits of South African languages resources (Moors et al., 2018; Sharma Grover et al., 2011), the SADiLaR^{xx}, masakhane^{xxi} and Lanafrica^{xxii} repositories, and the articles in topic conferences only (Joshi et al., 2021). Many advances for African languages do not reach the top conferences, however, be it due to travel and registration costs, target audience whom one wants to know about the work, and possibly also the perceived quality since efforts may be perceived pedestrian when compared to the level of English HLTs that most reviewers are familiar with rather than it being compared to the state of the art for the language under investigation. The HLT audits and the SADiLaR repository are limited to South African languages only and only those tools developed with SADiLaR funding, masakhane to corpora only, and lanafrica is collecting only since 2022 and does not include tools. That is: relying on only any one source or strategy for data collection to categorise languages on their tooling resourced-ness is going to be substantially incomplete and therewith distorted. Even if they were to be complete, it then

would need to be converted into a scale to measure, and it still would not indicate anything about the broader context and the features of funding, people, or education. Conversely, an expert-based classification using the aforementioned features remains opaque at least in part due to incomplete evidence in that case. Data for the features may be found to substantiate a categorisation value, but some of them have inherently fuzzy boundaries. For instance, evidence of teacher shortage can be obtained from a Department of Education, but a teacher shortage of 20% is different from a shortage where only 1 in 10 schools offer classes in the language to be categorised, but so are 25% and 1 in 9, respectively. It thus serves to refine some of the values for the features to enable a more robust categorisation.

Meanwhile, as a preliminary exercise, we use the expert-based approach. For the very LRL and official languages in South Africa, we classify Xitsonga, Venda, Sepedi, siSwati, isiNdebele, and Sotho as such. Other examples from Africa, among many, are Mboshi and Malagasy. The LRL category may be most interesting theoretically. IsiXhosa from South Africa clearly is solidly LRL. Concerning tooling but even more so political clout and *lingua franca* practices, one might argue for a sort of ‘upper LRL’, which are the “hopefuls scraping by” in Joshi et al.’s terminology, being isiZulu and Setswana in South Africa and Kiswahili in Tanzania and Kenya. They are still LRL but have a few more resources than the likes of isiXhosa. As RL, there are Afrikaans and South African English for South Africa, as well as Portuguese and Arabic in Africa, and, e.g., Italian internationally. Examples of HRL are Dutch, German, French, and Mandarin. The only language that is very HRL is English.

4. Discussion

The use of the term low-resourced in the literature has hitherto evidently been problematic. The views adduced from Mika Härmäläinen (2021), that in the NLP literature Chinese (1.2 billion speakers), Arabic (422 million speakers), Bengali (228 million speakers), Japanese (126 million speakers), Vietnamese (76 million speakers), Dutch (24 million speakers), Sinhala (17 million speakers) and Finnish (5 million speakers), have had several tasks referred to as low-resourced is, at best, based on a subjective personal scale or with English being the benchmark across this great divide. For the term low-resourced to carry any scientific currency in appreciation of research and application projects that avail of NLP tasks more generally, and indeed in other specialized environments, the exigency exists to define in a scalable way the range of language resources.

We have argued for eleven dimensions of resourced-ness in order to address the semantic obfuscation presented by the current use of the term low-resourced. The eleven dimensions allowed for the matching of resource characterization onto a 5-level categorization of resources toward a familiar 5-point Likert scale: VLRL; LRL, RL, HRL, VHRL as discussed in section 3.2.1. These categories provide an answer to the problem raised by Mika Hämäläinen (2021), that of lumping all the languages other than English together as LRL, something that he rightly observed as evidently incorrect. LRLs are clearly not a homogenous group, and the 5-level categorization provided here means that more languages can be allocated into these categories in a more meaningful way. This may also assist with “LRL sessions” at NLP conferences, to clarify what experiences and research are asked for, and for authors to position their contributions more precisely. Likewise, not all projects, be they in digital humanities or other fields, will have the same pace of progress due to the different amounts of NLP tasks entailed in the overall project, and the prerequisites they assume that have to be met before one can go start investigating research question from, say, sociology or history.

The categories as defined from the eleven dimensions are useful in providing a language policy and planning framework for the people responsible for the development of languages and language technologies that are in the levels VLRL and LRL. The deployment of resources and instruments required to develop them can henceforth be based on a scientific criterion, instead of either population demographics alone or discovered tool, corpus, or research paper-based counting. It has been shown that the fact that a language has fewer speakers (e.g., Finnish) or is an endangered language (e.g., Skolt Sami) (Hämäläinen, 2021) does not mean that it is a LRL. The categorization will therefore provide government and language policy experts a further mechanism in developing more informed language policies and more effective language implementation frameworks that address the characteristic elements of language category as articulated in the eleven dimensions. Such policies and frameworks may need to set feature cut-off points appropriate for their context, such as about the comparative number of language teachers in schools.

The categories also provide a rubric for planners to track the progress in the development of each language through the categories VLRL; LRL; RL; HRL and VHRL. This is particularly useful within the South African context, as the South Africa Constitution (1996) clearly articulates the commitment to develop all the eleven official languages in order to achieve “parity of esteem” between all the eleven official languages. The categorization can again be a useful framework to ascertain the categories of all the official languages, determine what tools or resources are required for each language and track the progression of the

languages from the low level to higher level of resourced-ness, and whether this has achieved the parity among all of them.

5. Conclusions

The paper proposed a list of features on the broader context of how a language is situated and, from that, a feature-based classification of resourced-ness of languages. It is based on 11 features and the classification was shaped in a 5-point scale, from very low to very high-resourced. Unlike a quantitative counting of number of speakers and number of corpora and tools, it positions the resourced-ness in a broader context that a language is positioned in.

The broader context of a language affects development of human language technologies in numerous ways and its characterisation may contribute to language policy development and implementation, as well as assessment of extant policies on language technology development. Further research might look into, inter alia, the types of resources that are available for each of South Africa's official languages and what low-hanging technology and policy development can be targeted for each according to the available resources to strategise to obtain the most realistic and effective output when augmenting the limited resources.

Acknowledgements

This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Number 120852).

References

- Berment, V.** 2004 Méthodes pour informatiser des langues et des groupes de langues peu dotées. PhD Thesis, J. Fourier University – Grenoble I, May 2004. <https://theses.hal.science/tel-00006313/document>
- Besacier, L., Barnard, E., Karpov, A., Schultz, T.** 2014 Automatic Speech Recognition for Under-Resourced Languages: A Survey. *Speech communication*, 56:85-100. <https://doi.org/10.1016/j.specom.2013.07.008>
- Bosch, S.E., Jones, J., Pretorius, L., Anderson, W.** 2007. Computational morphological analysers and machine readable lexicons for South African Bantu languages. *Localisation Focus – The International Journal of Localisation*, 6(1): 22-28.
- Moors, C., Calteaux, K., Wilken, I., Gumede, T.** 2018. Human Language Technology Audit 2017/8. In: *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists*. ACM. pp296-304.
- Cieri, C., Maxwell, M., Strassel, S., Tracey, J.** 2016. Selection Criteria for Low Resource Language Programs. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).
- Gerdes, K., Guillaume, B., Kahane, S., Perrier, G.** 2018 SUD or Surface-Syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. *Universal Dependencies Workshop 2018*, Nov 2018, Brussels, Belgium.
- Hämäläinen, M.** 2021. Endangered Languages are not Low-Resourced. *Preprints 2021*, 2021040113 (doi: 10.20944/preprints202104.0113.v1).
- Hammarström, H., Thornell, C., Petzell, M., Westerlund, T.** 2008. Bootstrapping language description: the case of Mpiemo (Bantu A, Central African Republic). In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association (ELRA).

- Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D.** 2021. A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568 June 6–11, 2021. Association for Computational Linguistics.
- Hocking, J., Puttkammer, M.** 2016. Optical Character Recognition for South African Languages. Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech 2016). *IEEE Xplore*, 10.1109/RoboMech.2016.7813139
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., Choudhury, M.** 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp6282–6293, Online. Association for Computational Linguistics.
- Sharma Grover, A., Van Huyssteen, G., Pretorius, M.** 2011. The South African Human Language Technology Audit. *Language Resources and Evaluation*, 45(3): 271–288.
- Keet, C.M., Khumalo, L.** 2014. Basics for a grammar engine to verbalize logical theories in isiZulu. In: *8th International Web Rule Symposium (RuleML'14)*, A. Bikakis et al. (Eds.). Springer LNCS vol. 8620, 216–225.
- Khumalo, L.** 2021. Corpus development and computational tools for the development of African languages. In Buthelezi, Z. and Khumalo, L.(Eds). *IsiZulu and Social Transformation in Southern Africa*. CAL Book Series: University of Zululand.
- Kirmizialtin, S., Wrisley, D.J.** 2022. Automated transcription of non-Latin script periodicals: a case study in the Ottoman Turkish Print Archive. *Digital Humanities Quarterly*, 16(2): 16.2.
- Krauwier, S.** 2003. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap. In: *Proceedings of the 2003 International Workshop Speech and Computer SPECOM-2003*, Moscow, Russia, 2003, pp. 8–15.
- Magueresse, A., Carles, V., Heetderks, E.** 2020. Low-resource Languages: A Review of Past Work and Future Challenges. *Arxiv technical report* <https://arxiv.org/abs/2006.07264>
- Marneffe, M.C. de, Manning, C.D., Nivre, J., Zeman, D.** 2021. Universal dependencies. *Computational Linguistics*, 47(2): 255–308.
- Ndaba, B., Suleman, H., Keet, C.M., Khumalo, L.** 2016. The effects of a corpus on isizulu spellcheckers based on n-grams. In: *IST-Africa 2016*. IIMC International Information Management Corporation. 11–13 May, 2016, Durban, South Africa.
- Park, H.H., Schwartz, L., Tyers, F.M.** 2021. Expanding universal dependencies for polysynthetic languages: A case of St. Lawrence island Yupik. In: *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. pp131–142. Association for Computational Linguistics (2021).
- Pauw, G. de, Wagacha, P.W.** 2007. Bootstrapping morphological analysis of g̃ik̃uỹu using unsupervised maximum entropy learning. In: *Proceedings of the Eighth Annual Conference of the International Speech Communication Association (INTERSPEECH'07)*, pp1517–1520. ISCA.
- Pretorius, L., Bosch, S.E.** 2003. Computational aids for Zulu natural language processing. *Southern African Linguistics and Applied Language Studies*, 21(4): 265–282.
- Ranathunga, S., Lee, E-S. A., Skenduli, M.P., Shekar, R., Alam, M., Kaur, R.** 2021. Neural Machine Translation for Low-Resource Languages: A Survey. *Arxiv Technical Report*, <https://arxiv.org/pdf/2106.15115.pdf> 29 June 2021.
- Schryver, G.-M. de, Prinsloo, D.J.** 2004. Spellcheckers for the South African languages, Part 1: the status quo and options for improvement. *South African Journal for African Languages*, 1: 57–82.
- Steinberger, R., Ombuya, S., Kabadjov, M., Pouliquen, B., Della Rocca, L., Belyaeva, J., de Paola, M., Ignat, C., van der Goot, E.** 2011. Expanding a multilingual media monitoring and information extraction tool to a new language: Swahili. *Language Resources & Evaluation*, 45:311–330.
- The Constitution of the Republic of South Africa.** 1996. <https://www.gov.za/documents/constitution/constitution-republic-south-africa-1996-1>
- Tsvetkov, Y.** 2017. Opportunities and challenges in working with low-resource languages. Carnegie Mellon University. <https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf>
- Xu, D.** 2023. Digital Approaches to understanding Dongba pictograms. *International Journal of Digital humanities*, 4: 131–146.

ⁱ <http://lloydbleekcollection.cs.uct.ac.za/>

ⁱⁱ From a Medium article of 2022 by Felix Naumann <https://medium.com/neuralspace/low-resource-language-what-does-it-mean-d067ec85dea5> (last accessed in October 2022), which was likely lifted from Tsvetko's lecture slides of 2017 <https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf> (slide 26).

ⁱⁱⁱ <https://arxiv.org/abs/2006.07264>

^{iv} Good OCR software does not only try to recognize the characters, but also add probabilities to words and characters that it is not sure of. This may be ‘fixing’ a string so that it passes a spelling check for the language it is optimized for (e.g., changing the isiZulu *wena* into the English ‘wend’) up to analysis the sentence to choose among alternative corrections (e.g., if the sentence misses a verb, then the candidate fixes for the unclear string will select a verb).

^v The three examples were not merely theoretical, but they have occurred to the authors.

^{vi} Besides the Medium article (fn 1), see also the “High resource languages vs low resource languages” by Emily Bender in *The Gradient*, 14 Sept 2019, <https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/> (last accessed on 3-1-2023). It only characterizes high resource languages, leaving the ‘low’ implicit as not having all that.

^{vii} <http://www.blark.org/>

^{viii} <https://microsoft.github.io/linguisticdiversity/assets/lang2tax.txt>

^{ix} The list of detected papers has not been made available by Joshi et al., but by their methodology, it would have missed at least the LREC2018 paper on Mboshi, being Rialland A, Adda-Decker M, Kouarata G-N, Adda G, Besacier L, et al. Parallel Corpora in Mboshi (Bantu C25, Congo-Brazzaville). 11th Language Resources and Evaluation Conference (LREC 2018), ELRA, May 2018, Miyazaki, Japan.

^x <http://extensions.openoffice.org/en/project/zulu-spell-checker>

^{xi} <https://sadilar.org/index.php/en/applications/spelling-checkers-for-south-african-languages>

^{xii}

https://www.uct.ac.za/sites/default/files/content_migration/uct_ac_za/49/files/2022_HUM_UG_Handbook.pdf

^{xiii} For instance, the isiZulu majors at the University of KwaZulu-Natal

<https://soa.ukzn.ac.za/clustersdisciplines/language-and-literature/isizulu/>

^{xiv} <https://www.tcd.ie/courses/undergraduate/courses/irish-jh/> and

<https://www.tcd.ie/courses/undergraduate/courses/early-and-modern-irish/>

^{xv} <https://www.unibz.it/en/services/language-centre/> it offers Arabic, Chinese, English, French, German, Italian, Latin, Russian, and Spanish.

^{xvi} <https://www.wur.nl/en/education-programmes/wageningen-into-languages/language-courses.htm> It offers Arabic, Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Norwegian, Portuguese, Russian, Swedish, Spanish, and Turkish.

^{xvii} <https://uctlanguagecentre.com/>

^{xviii} <https://languagecentre.sun.ac.za/>

^{xix} <https://www.usaf.ac.za/how-language-scholars-and-activists-are-keeping-multilingualism-on-universities-transformation-agenda/>.

^{xx} <https://repo.sadilar.org/>

^{xxi} <https://www.masakhane.io/>

^{xxii} <https://lanfrica.com/>