Tackling Language Modelling Bias in Support of Linguistic Diversity

Bella, G.; Helm, P.; Koch, G.; Giunchiglia, F.

DOI
10.1145/3630106.3658925

Publication date
2024

Document Version
Final published version

Published in
ACM FAccT '24

Tackling Language Modelling Bias in Support of Linguistic Diversity

Gábor Bella
gabor.bella@imt-atlantique.fr
IMT Atlantique, Lab-STICC, UMR CNRS 6285
Brest, France

Paula Helm
p.m.helm@uva.nl
University of Amsterdam, Faculty of Humanities
Amsterdam, The Netherlands

Gertraud Koch
gertraud.koch@uni-hamburg.de
University of Hamburg, Institute of Anthropological Studies on Culture and History
Hamburg, Germany

Fausto Giunchiglia
fausto.giunchiglia@unitn.it
University of Trento,
Dept. of Information Engineering and Computer Science
Trento, Italy

ABSTRACT

Current AI-based language technologies—language models, machine translation systems, multilingual dictionaries and corpora—are known to focus on the world's 2–3% most widely spoken languages. Research efforts of the past decade have attempted to expand this coverage to 'under-resourced languages.' The goal of our paper is to bring attention to a corollary phenomenon that we call *language modelling bias*: multilingual language processing systems often exhibit a hardwired, yet usually involuntary and hidden representational preference towards certain languages. We define language modelling bias as uneven per-language performance under similar test conditions. We show that bias stems not only from technology but also from ethically problematic research and development methodologies that disregard the needs of language communities. Moving towards diversity-aware alternatives, we present an initiative that aims at reducing language modelling bias within lexical resources through both technology design and methodology, based on an eye-level collaboration with local communities.

KEYWORDS

language modeling bias, linguistic diversity, low-resource languages, natural language processing, value-sensitive design

ACM Reference Format:
Gábor Bella, Paula Helm, Gertraud Koch, and Fausto Giunchiglia. 2024. Tackling Language Modelling Bias in Support of Linguistic Diversity. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658925

1 INTRODUCTION

The notion of *digital language divide* refers to the gap between languages with and without a considerable representation on the Web and within the worldwide digital infrastructure [46]. As shown by [37] about 10 years ago, less than 5% of the world's 7–8,000 languages have an even remotely significant representation on the Internet. The same orders of magnitude remain valid today, despite the progresses of a decade [33]. Due to the inextricable link between language, culture, and society (as we show through many examples in this paper), the ability of persons and peoples to express themselves in their own language, dialect, or even sociolect,[1] is determinant in maintaining their identity and their unique perspective in which ideas and worldviews are anchored, and which are thus crucial for the dignity of human beings as well as from the point of view of epistemic justice [29, 45, 52]. In the field of language technology and research, riding the wave of the recent breakthrough of neural AI, the last decade saw a surge in multilingual language tools and resources for 'under-resourced languages.' The promise of neural language technology is its apparent language-agnosticism: when fed with raw corpora that are large enough, statistical and neural computation makes language processing abilities emerge in an inductive manner, seemingly independently of the underlying language structures.

Despite its undeniable results over hundreds of languages, the linguistics-unaware *modus operandi* of neural language research has been criticised from multiple perspectives. From a methodological point of view, due to an insufficient understanding of researchers about the corpora, the languages and, ultimately, the cultures being worked upon, major quality problems in research output remain hidden behind precision–recall figures and eventually go unnoticed by the scientific community [38]. Ethics-wise, the attitude of first-world experts who pretend to 'save the day' in the Global South by applying blanket solutions to languages with which they have no contact or understanding has been pointed out as fundamentally neocolonial [11, 50].

With the goal of strengthening the quantitative backing of such criticisms, our paper draws attention to the phenomenon that language technologies may exhibit unintended preferences towards certain linguistic and semantic constructs, leading to a performance imbalance across languages even in the case of comparable data sizes and parameters. Accordingly, the first of the three contributions of our paper is a formal, quantitative definition of this

[1] For simplicity, in the rest of the paper we will use the term *language* in a broad manner in order to encompass dialects and sociolects.

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil

ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658925

phenomenon, that we refer to as *language modelling bias*.[2] Our definition is based on an abstract, task-dependent interpretation of *performance*, allowing it to be applied to neural, statistical, or knowledge-driven language technologies. This form of bias that stems from unequal AI performance is also understood as an AI trustworthiness issue [39].

Language modelling bias is tightly related to a second key notion of our paper, *linguistic diversity*, that refers to linguistic features and ideas that are 'hard to translate' across languages. Our second contribution is to show how, alongside its technological origins, language modelling bias is also caused by flawed methodologies and a lack of in-depth ethical reflection about language technology development. Consequently, we argue that, in order for technology design to do justice to linguistic diversity, an engaged participation of local communities is required, via a local institutional framework and rigorous co-design.

As our third contribution, we present the case study of reducing bias within the *Universal Knowledge Core* lexico-semantic database, a large-scale multilingual lexical resource, via the language-community-driven *LiveLanguage* initiative. While other efforts in this direction exist (see i.e. the Masakhane initiative [44] and earlier work dealing with typological diversity in NLP [31]), our approach places the needs of local communities at the centre, an important aspect that has so far been marginalized. Finally, as the uneven digital representation of languages is a complex problem set, of which linguistic diversity and bias are but a puzzle piece, our solutions are necessary, yet not sufficient conditions for bridging the digital language divide.

The rest of the paper is organised as follows. Section 2 defines and discusses the notions of linguistic diversity and bias. Section 3 provides examples of bias from AI-based language technology. Section 4 provides a critique of methodologies in language technology research and development, and proposes alternatives. Sections 5 and 6 describe the technical and methodological aspects of the LiveLanguage initiative, aiming to put into practice the principles discussed in Sections 2–4. Finally, Section 7 situates the state of our work with respect to long-term goals.

2 LINGUISTIC DIVERSITY AND LANGUAGE MODELLING BIAS

The term *linguistic diversity* has a positive connotation: evocative of *biodiversity*, its association to language implies the preservation of the variedness of the world's linguistic landscape. Although our own point of departure is one of preserving diversity, we are wary of naïvely celebrating it without a proper conceptualization Therefore, we differentiate between an understanding of linguistic diversity as a *descriptive* and as a *normative* concept [30]. The former points to the actual notions of difference that underlie our understanding of diversity, both in the field of linguistics [25] and as a design strategy in computational systems [23, 48]. In what follows, we will focus on the latter normative conception of diversity, i.e. the values we associate with it as the objective of our work.

[2]Thus, we understand *language modelling* in a broad sense to cover any form of algorithmic model of any aspect of language, as opposed to the very narrow sense in which recent AI research understands pretrained neural *(large) language models*.

Diversity can be seen as a value that is either intrinsic or instrumental [57]. In the intrinsic version, diversity is good by and for itself, and evokes associations with pluralism, tradition, and authenticity [18, 54]. Following a more instrumental stance, the UNESCO 'Convention on the Protection and Promotion of the Diversity of Cultural Expressions' supports the idea of relating the preservation of linguistic diversity to values of tolerance, inclusion and dignity [52].

Following such instrumental normative understanding of diversity, in cross-lingual and cross-cultural contexts it is also important to acknowledge a necessary compromise between linguistic unity, i.e. the effectiveness of communication, and diversity. Trade languages such as English or Swahili, for example, spoken by various peoples across the globe, enable mutual understanding and exchange of ideas. Also, diversity should not be treated as a commodity that can be exploited, e.g. as part of corporate PR-washing strategies [56]. Based on these consideration, we settle on an understanding of diversity as an instrumental value. As such, simply preserving or promoting linguistic diversity through technological representation is not yet sufficient as a normative goal, it is also the means that are critical as well as the kinds of implementations that are enabled through this activity.

Given these conceptual clarifications, we embrace linguistic diversity as an objective, together with the idea that computational efforts can be instrumental to achieve it. In this perspective, we first provide a general and descriptive definition for linguistic diversity, followed by an operational and normative interpretation of what dealing with linguistic diversity implies in practice, also in terms of objectives to be achieved.

> **Linguistic diversity** is observed across two (or more) languages if one language possesses a particular linguistic construct through which it can express an idea concisely, while the same construct is absent from the other language that, in consequence, needs to express it through different constructs, if at all.

This definition is general, with the term *linguistic construct* possibly referring to any lexical, syntactic, morphological, etc. phenomenon; yet, the reference to *expressing an idea* implies that the construct in question *determines the meaning* of the utterance in which it is used. For speakers, the ideas expressed by such constructs are often inextricably embedded in the local geographical and cultural context. We illustrate our point with two such constructs: *lexical untranslatability* and *inalienable possession*.

Lexical untranslatability refers to language-specific terms that do not have equivalents across languages. Linguists and ethnographers have for long studied such phenomena, notably in the fields of *colour terms*, *geography*, *body parts*, or *kinship*, the last one being perhaps the most thoroughly studied [42]. If for siblings, for example, English only distinguishes by gender (*brother*, *sister*), many other languages take into account the relative age of the sibling and the gender of the speaker as well. Thus, the Maori word *teina* means *elder brother* if it is pronounced by a male speaker, and *elder sister* if it is pronounced by a female. Likewise, while English has a single term for *cousin*, a speaker in South India respectful of culture will choose out of 16 possible cousin-type terms, depending on gender,

age, the mother's or father's side, etc. In the Kaxinawa language from Amazonia, the broad kin term *siu'i* refers to people who are already known to the speaker's community, also implying a possible blood relation. Such radically different organisations of kinship terminology reveal an underlying diversity in family structure and social organisation.

Inalienable possession stems from the boundary of semantics, syntax, and morphology. It is widely present in Native American and Australasian languages, where abstract—yet for us natural—concepts such as *mother* or *head* (as a body part) cannot be expressed as single words (free morphemes), but only together with their possessor (i.e. as the combination of two bound morphemes): *my mother*, *your head*.

Despite the rich literature in linguistics of such phenomena, they are for the most part neglected in the AI and computational linguistics communities. Tools such as machine translators or multilingual lexicons are rarely, if at all, evaluated with respect to their support of linguistic diversity. In this paper, we attempt to approach this problem through the notion of *language modelling bias*.

In the context of AI language technology, the notion of *bias* has so far been used to refer to patterns of stereotypes and preferences towards social groups, most often concerning learning-based language processing systems [13]. In terms of social groups, studies have mostly focused on gender, ethnicity, and race, but also other forms of bias (religion-related, age-related, political, socioeconomic, etc.) [21]. To our knowledge, the term *language modelling bias* has not been used so far in any way similar to ours. Many of the underlying exploitative mechanisms have, however, been pointed out, in particular in relation to the most disempowered social groups, namely small indigenous speaker communities [12, 50]. In terms of actual bias in AI systems and data, the research closest to ours concerns *inductive bias* in language models towards certain morphological and syntactic structures [47, 55]. We present these works more in detail in Section 3. In [9], Bender et al. show examples of language technology that do not hold their promise with respect to language-agnosticism. In [14], Blodgett et al. study the (non-)representation of the vernaculars of social groups within language resources. They point out that English linguistic corpora tend to exclude the register of speech used by African-Americans, the non-representation of which causes a bias in the abilities of the AI systems trained on top of them. We identify this as a particular case of language modelling bias, even if in the paper cited it is (correctly) also framed as a form of racial bias across sociolects of a single language. As stated in the introduction, our notion of language modelling bias aims to encompass sociolects as well as dialects, as the effect of bias on these different categories of speaker groups cannot be distinguished.

Intuitively, language modelling bias is observed in language technology when, *due to its design*, a system represents, interprets, or processes utterances in certain languages less precisely or less efficiently than in others, thereby negatively affecting the communication ability of speakers of that language. More formally:

> A technology t that supports the languages (dialects, sociolects) $L = \{l_1, ..., l_N\}$ has **language modelling bias** if there exist a pair of languages $l_A \in L, l_B \in L$, an operation o_t performed by t, a set of utterances U_A in language l_A given as input to o_t, and a set of *analogous* utterances U_B in language l_B, such that the performance of o_t over U_A is distinctly better than its performance over U_B: $\text{Perf}(o_t(U_A)) >> \text{Perf}(o_t(U_B))$.

In order to obtain a quantitative measure of language modelling bias b, we use the *coefficient of variation* over the performance values measured across N languages:

$$b_t = \frac{\sigma_t}{\mu_t} = \frac{\sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\text{Perf}(o_t(U_i))^2 - \overline{\text{Perf}}^2)}}{\overline{\text{Perf}}}$$

where $\overline{\text{Perf}}$ is the mean performance of t over the languages. The intuition behind this formula is that the more varied the performance of the operation o_t over the set of languages L, the higher the standard deviation σ_t. In order to compute the language modelling bias b_t, we normalise the standard deviation of the performances by their mean, thereby obtaining a performance-metric-independent and more cross-comparable measure.

It is important to note that we use the term *performance* in an abstract manner, as it can be measured in many ways depending on the task being evaluated. For example, performance may refer to common metrics such as precision, recall, BLEU, or ROUGE that all vary between 0 and 1, but it may also be based on perplexity (e.g. to evaluate pretrained language models) or semantic distance (e.g. to evaluate semantic correctness of machine translation) where the values have no upper bound and the lower they are the better.[3]

If applied unconditionally, the definition above will find bias everywhere, as in practice no technology performs equally well on any two languages. As most language processing systems today are data-driven, any difference in training data size or quality will inevitably lead to uneven performance. While it may be revealing to evaluate bias in existing, pre-trained systems, our aim is to use language modelling bias to examine deeper, structural limitations built into language processing algorithms, representational models, resources, or methodologies. In practice, however, from merely observing the output of a system, it may be difficult to understand whether lower performance is caused by its structural properties or by contingent factors such as resource completeness or training data size. The distinction is important as the latter kinds of issues can in theory be mitigated by adding more (or better-distributed, higher-quality, etc.) data to the system, while structurally determined bias can only be addressed by redesigning the technology itself (as well as the methodologies that may have caused the bias). In order to focus on the structural sources of bias, a careful selection of the set of input utterances U may be necessary, and the systems typically need to be retrained or repopulated with corpora balanced across languages.

We recognise that bias is generally unavoidable. Contrary to a blanket critique of bias as a phenomenon in itself, it is now accepted in humanities and social science research that all knowledge, all insights, and even all data are situated, i.e. they always reflect a particular point of view in space and time that is influenced by culture, history, politics, economics, epistemology, and so on [22, 27].

[3] In order for the coefficient of variation to be interpretable, however, we do take the reasonable assumption that Perf is a ratio scale.

Unbiasedness is therefore a deceptive goal that, instead of solving social problems, reproduces problematic ideas, such as the unrealistic imaginary that technology can be neutral [5]. It is therefore important to be upfront about when and for what reasons a certain bias is problematic and needs to be addressed, and that this does not lead to an elimination of bias, but to a different, more transparent and just bias [28]. This is the case, for example, when bias targets already vulnerable, underrepresented, and marginalized groups. In our case, the social group in question is clearly the community of speakers of a given language, however heterogeneous it may be otherwise (according to social status, culture, gender, race, ethnicity, religion, etc.). Being the native or second-language speaker of a language in itself determines one's access to information, and the language technology that enables this access affects one's ability to communicate, on the Web or elsewhere.

3 EXAMPLES OF LANGUAGE MODELLING BIAS

This section presents examples of language modelling bias in mainstream AI language technology: within the structure of multilingual lexical databases, within neural language models, and finally various manifestations of language modelling bias in machine translation systems.

Bias in Lexical Databases. As a generalisation of bilingual dictionaries, the 2000s saw the appearance of *multilingual lexical databases* that map words, based on their meanings, across large numbers of languages. As shown in the survey [24], several of these multilingual databases interconnect words from hundreds of languages, mapping the words of each language to the 100 thousand English word meanings (so-called *synsets*) of Princeton WordNet [41]. On the one hand, this choice makes practical sense, as among all similar resources, WordNet offers by far the widest coverage of word meanings. On the other hand, it results in a strong bias towards the English language and Anglo-Saxon culture in general, as the expressivity of the database is limited to notions for which a word exists in English [6, 24]. Figure 1 provides a simple example from the food domain, known to be culturally, and thus also linguistically, diverse. It shows how a biased lexical database maps together words in Swahili and Japanese meaning *uncooked rice*, *cooked rice*, and *brown rice*. The degree of information loss is flagrant: while both Swahili and Japanese provide fine-grained lexicalisations about the various forms of rice, the many-to-many mapping that results from passing through English masks all fine-grained differences, resulting in both a loss of detail and incorrect translations when one moves from Swahili to Japanese or vice versa. The diversity-diminishing bias towards the English language and Anglo-Saxon cultures is also found in other domains that are well-known to be diverse across languages: family relationships, school systems, etc.

Applying the definition from Section 2, we compute the language modelling bias of the lexical models of four multilingual lexical databases: the first and second versions of the Open Multilingual Wordnet (OMW, OMW2) [15, 16], IndoWordNet (IWN) [10], and BabelNet (BN) [43]. The figures were not computed from the actual contents of the databases—that are necessarily incomplete and thus are not representative of their structural properties—but rather from a gold-standard cross-lingual mapping dataset [24], covering a diverse set of nine languages from five phyla. Coverages are theoretical in the sense that they mean the percentage of cross-lingual mapping relationships that each model is *structurally able to represent* with respect to the gold standard mappings. The dataset contains three mapping relation types: *equivalent meaning*, *broader/narrower meaning*, and *lexical untranslatability*. We evaluate bias as follows:

- technology t: multilingual lexical databases {OMW, OMW2, IWN, BN};
- operation o_t: cross-lingual mapping of word meanings;
- languages L: {English, French, Italian, Chinese, Hindi, Tamil, Malayalam, Hungarian, Mongolian};
- utterances U: 32 concepts from six linguistically diverse domains, lexicalised by the languages above through 160 words and 128 *interlingual gaps* representing lexical untranslatability;
- performance Perf: defined as the coverage (recall) of gold-standard cross-lingual mappings that the lexical database is able to express.

Figure 2 shows performance figures. Bias is not concerned with absolute coverage values, but rather with how coverage varies across languages. OMW shows a marked bias towards European languages and English in particular (68% coverage) while Asian languages are mapped suboptimally (49–51%), with a bias of $b_{\text{OMW}} = 0.120$. This is explained by the fact that pivot concepts in OMW are limited to the meanings of English words. IWN, where the pivot is Hindi, unsurprisingly displays a bias towards Indian languages (75–76%) with other languages falling into the 59–68% range, with $b_{\text{IWN}} = 0.079$. BN and OMW2, on the other hand, are less biased due to the fact that their pivot concepts are not tied to any particular language. The bias of BN, $b_{\text{BN}} = 0.031$, the mapping coverage of which varies between 77% and 83%, is due to its lack of support for untranslatability relations. In the case of OMW2 (coverage 86–92%), the so far smallest bias $b_{\text{OMW2}} = 0.023$ is caused by limited expressivity in cross-lingual broader/narrower mappings.

Bias in Neural Language Models. Do neural language models favour, in terms of better performance, certain types of languages based on their grammatical features (e.g. word order, morphology)? This question has been discussed in previous work, also with respect to cross-lingual variations on specific grammatical features such as morphology [31, 40] or word order [55]. A growing number of studies exist on the effect of morphology on the results of popular language-agnostic tokenisation methods such as Byte-Pair Encoding: as tokenisation is a crucial preprocessing task, it affects most aspects of language model performance [4, 49].

Some works approach the evaluation of structural bias through abstraction from variations in training data and other contingent factors, in line with our position in Section 2. [47] and [55] both construct artificial languages that differ by single typological features in order to compare LSTM and Transformer architectures over their support of grammatical features. [31], on the other hand, evaluate five pre-trained large language models with respect to their accuracy in answering prompt-based questions. As an illustration, using our formalism, we compute the bias of the five models based on detailed evaluation data from [31, Table 3] as follows:

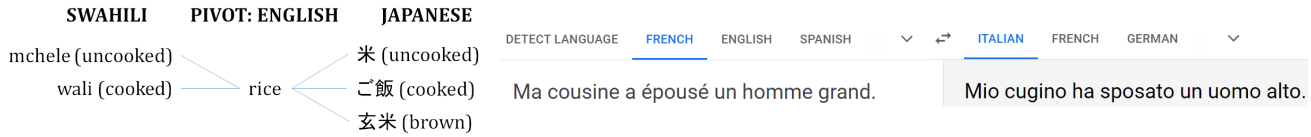
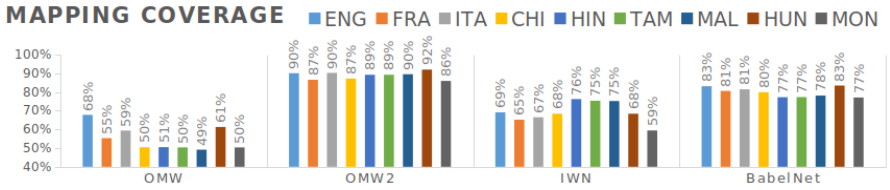


Figure 1: Left: biased cross-lingual mapping of words about various forms of 'rice' from a popular multilingual lexical database. Right: an example of language modelling bias in machine translation.



BIAS

Lexical DB	Bias
OMW	0.120
OMW2	0.023
IWN	0.079
BN	0.031

Figure 2: Bias in the expressivity of cross-lingual mappings of multilingual lexical databases. Left: per-language coverage (performance) of mappings for each database. Right: bias of each database.

- technology t: a language model from {mT5-XXL, PaLM-S, PaLM-M, PaLM-L, PaLM-2};
- operation o_t: LLM-based question answering;
- languages L: {English, Spanish, Italian, French, German, Swedish, Finnish, Slovak, Russian, Chinese, Swahili, Arabic};
- utterances U: 10k template-generated context–question–answer prompts per language;
- performance Perf: accuracy.

The bias values obtained are: $b_{\text{mT5}} = 0.22$, $b_{\text{PaLM-S}} = 0.31$, $b_{\text{PaLM-M}} = 0.23$, $b_{\text{PaLM-L}} = 0.08$, $b_{\text{PaLM-2}} = 0.05$. In this study, the largest and most performant model, PaLM-2, also proves to be the least biased. Here, actual models of varying sizes are being compared, thus the bias computed is largely dependent on training data and does not measure the structural properties of models.

Bias in Machine Translation. Machine translation (MT) has been a flagship task of AI-based language technology. Without claiming to be exhaustive, we point out three aspects of current MT technologies where linguistic bias can be observed: the non-handling of untranslatability, the variedness of vocabulary and grammar, and the use of a pivot language. Today's top MT systems, such as DeepL and Google Translate, make systematic mistakes over untranslatable terms, betraying the fact that this phenomenon is not specifically addressed by these tools.

For instance, when translating the English sentence *My brother is three years younger than me* to Hungarian, Korean, Japanese, or Mongolian, syntactically correct yet semantically absurd results are obtained [36].[4] These languages either have no equivalent word for *brother* or, when they do, the equivalent word is relatively rare (as *fiútestvér* in Hungarian). Based on training corpus frequencies, the MT system ends up choosing a semantically unsuitable word, such as *bátyám* meaning *my elder brother*, resulting in *My elder brother is three years younger than me.* A similar example, based on the example of *rice* from earlier in this section, is the English sentence *'This rice is tasty,'* machine-translated into Swahili as *'Mchele huu ni kitamu,'* meaning *'This raw rice is tasty.'* These are not cherry-picked exceptions but rather are examples of systematic mistakes from domains of high linguistic diversity.

Given the nature of the errors above—the use of words with incorrect meanings within otherwise syntactically correct sentences—and reusing results from [36], we propose a measurement of bias based on lexical semantics. We use as a measurement of performance the average semantic distance, more precisely the *least common subsumer distance*, between the meaning of each translated word and the expected gold-standard meaning, measured over the interlingual concept hierarchy published in [36].

- technology t: Google Translate {GT};
- operation o_t: translation from English;
- languages L: Russian, Japanese, Korean, Hungarian, Mongolian;
- utterances U: 50 English sentences from the British National Corpus containing kinship terms, from [36];
- performance Perf: average semantic distance (the lower the better) between the meanings of translated words and correct gold-standard concepts.

The average semantic distances obtained, as reported in [36], are $\text{Perf}(\text{GT}_{\text{ENG–RUS}}) = 0.34$, $\text{Perf}(\text{GT}_{\text{ENG–JAP}}) = 0.38$, $\text{Perf}(\text{GT}_{\text{ENG–KOR}}) = 0.90$, $\text{Perf}(\text{GT}_{\text{ENG–HUN}}) = 1.06$, $\text{Perf}(\text{GT}_{\text{ENG–MON}}) = 1.12$, which provides an overall bias of $b_{\text{GT}} = 0.49$ (to be compared with the biases of other MT systems).

A second form of bias concerns the variedness of vocabulary and grammar in MT output. In [53], Vanmassenhove et al. quantitatively compare the lexical and grammatical diversity between original and machine-translated text. Their definitions of diversity and bias are different from ours: by diversity they refer to the richness of the vocabulary and the complexity of the grammar of a document (normalised by document size and computed according to multiple grammatical constructs), while by bias they understand an uncontrolled loss of diversity due to MT. Still, their results are relevant for our argument: in [53], they report that, for the same language,

[4] Recently, these tools have added correct translations as second or third alternatives.

morphology in translated text becomes poorer with respect to original (untranslated) corpora, i.e. features of number or gender for nouns tend to decrease. This phenomenon affects morphologically rich languages in particular.

A third form of bias in MT systems is their use of English as a pivot language when translating between non-English language pairs. This practice is explained by the relative scarcity of bilingual training corpora for such language pairs, as well as scalability: the use of a pivot language reduces the need for trained models from $\binom{N}{2}$ to $N-1$, where N is the number of languages. Figure 1 (right) shows the case of French-to-Italian translation of a sentence meaning *my (female) cousin married a tall man.* While French and Italian use different words for male and female cousins (*cousin/cousine*, *cugino/cugina*), English does not. The result is that the gender of the cousin is 'lost in translation' and, as a form of combined linguistic and gender bias, it appears as a male in the translated text.

4 METHODOLOGY AS A SOURCE OF BIAS

In our view, bias in language technology is also due to methodological flaws in computational linguistics research and development practices. In Computational Linguistics, English has not only been the lingua franca of scientific communication, but also the de facto standard subject matter of research. [50] reports that between 2013 and 2021, 83% of papers accepted at the flagship ACL conference were explicitly or implicitly about English and 97% were about Indo-European languages. The 2010s saw an emerging interest in multilingual language technology, and of a new research sub-field targeting 'low-resource' (or 'under-resourced') languages, previously neglected by mainstream research. The recent progress made in supporting new languages is undeniable—for example, as of early 2024, Google Translate supports 133 languages, while Meta claims to have broken the 200-language barrier with its (hyperbole ahead) *No Language Left Behind* (NLLB) machine translation project [51]. Nevertheless, in line with the 'zero-shot' data-driven ethos [11] of deep neural AI, mainstream low-resource language research aims to provide a solution for multiple, preferably tens or hundreds of languages at the same time, shunning human involvement (linguists, field workers, native speakers, final users) in the name of cost efficiency.

In many cases, the languages being addressed are not understood by the people working on them, who are therefore not able to judge the quality of the data and algorithms on which they are relying. This leads to methodological errors that remain hidden within systems, and of which the speaker communities only see the negative consequences in terms of the low-quality tools they are offered. Such errors may appear in multiple development steps, as in our three examples below on (1) corpus generation, (2) corpus preprocessing, and (3) evaluation.

(1) In the context of corpus generation, [38] report that researchers are not always familiar with the corpora they are using. For instance, when Wikipedia is scraped automatically, the contents of pages can be of low quality due to the use of machine translation, or may not even correspond to the language by which they are tagged. This results in systems trained and evaluated on low-quality text or even on the wrong language. (2) In corpus preprocessing, the practice of 'removing accents and special characters' (that are only special to engineers unaware of their role in languages other than theirs), that has become commonplace for English, has a negative effect on languages where these characters play an important linguistic role, e.g. in the disambiguation of meaning. An indiscriminate removal of 'special' characters results in bias against these languages. (3) In the context of evaluating a tool or resource, it is important to understand the limitations of evaluation metrics with respect to what they are or are not able to measure. For example, in machine translation, the standard BLEU metric is known not to measure the semantic similarity of the reference and the automated translations, while the METEOR metric takes synonymy into account, although only for English [2]. None of these metrics are reliable in the presence of lexical diversity, as exemplified in section 3, where cross-lingual hypernymy can be the preferred method of selecting the best possible translation candidate [36].

The methodology and the goals of such research raise a multitude of concerns that are both ethical and scientific. Linguists such as Bird [11] claim mainstream Western 'low-resource language' research to be postcolonial, with Western researchers unilaterally setting developmental goals and providing technological solutions to reach them. Most often, native speakers are not involved in the process, or when they are, they play subordinate roles such as annotator or validator. These criticisms are in line with what Irani et al. describe as 'postcolonial computing,' and with the four-dimensional ways forward that the respective authors propose [32].

Bird [11], Schwartz [50], or the researchers of the successful *Masakhane* project on African languages [44], have been advocating alternative, 'decolonising' approaches to multilingual research in AI and to working specifically with indigenous linguistic communities, based on an understanding of power imbalance and the difference in epistemologies between the researcher and the local community, and overcoming them through deliberate effort. Bird emphasises *co-design* of technology with communities, based on their perceived goals and needs. He observes the importance of *vehicular* or *trade languages* in addressing local vernaculars—beyond Spanish, French, or English, also Arabic, Persian, Hindi, Urdu, Amharic, Hausa, or Swahili are also widely used trade languages. In [12], a *multipolar model* is proposed for working with language communities, where trade languages function as bridges or pivots across local languages and vernaculars. Along a similar philosophy, *Masakhane* adopts a research methodology they call *participatory*, which makes sure that human agents are from local communities or, if this is not entirely possible, at least knowledge transfer takes place [44]. Human-based evaluation is emphasised in addition to conventional automated methods that are, justly, deemed inefficient in low-resource scenarios.

We endorse the idea of a co-design methodology where local communities exercise decisional power and property rights over research outcomes. We also embrace Bird's multipolar model, both as a methodological approach and as a high-level system architecture for the development of linguistic knowledge. Yet, we warn against the potential bias inherent in hub-spoke architectures in favour of the hub, as shown in section 3. This bias is avoidable through appropriate design, as we show in Section 5. With respect to the communities targeted, we present a different perspective: while these authors adopt the viewpoint of small and disempowered indigenous communities (e.g. Australian aboriginals, Native

Language	L1+2 speakers	Rank	Articles	Language	L1+2 speakers	Rank	Articles
English	>1B	1	6,624,314	Swahili	80M	83	76,417
Indonesian	300M	22	639,717	Hausa	77M	123	21,190
Bengali	300M	63	134,966	Pashto	40M	127	17,202
Marathi	100M	74	90,421	Scottish Gaelic	50k	133	15,859
Breton	200k	82	78,361				

Table 1: Contrast of the number of speakers (as 1st or 2nd language) and the number of Wikipedia articles for a selection of languages

Americans), we point out the need for a finer-grained typology, in order to develop an ethical framework that best corresponds to the community at hand, sometimes markedly different from small indigenous groups. For instance, [37] and [33] divide languages into five and six clusters, respectively, according to their online support. Communities with tens of millions of speakers are rarely disempowered linguistic minorities. As shown in Table 1,[5] languages such as Bengali, Urdu, or Indonesian are each spoken by 100 million people or more. Swahili, Hausa, or Pashto are each spoken by at least 50 million. Yet, the online presence of these languages is nowhere representative of such numbers.[6]

Languages such as Breton or Scottish Gaelic fall into yet another category, that of endangered minority languages of the Global North. These languages are characterised by a small number of speakers in steep decline, yet with an economic and socio-cultural support much higher than that of indigenous minorities in other parts of the world. This is also reflected in Table 1 where, in terms of Wikipedia content, Breton (200 thousand speakers) is on a par with Swahili (80 million) and Scottish Gaelic (50 thousand) with Pashto (at least 40 million).

Compared to the small indigenous communities targeted by Schwartz and Bird, these languages enjoy a non-negligible level of institutional backing: official language status and administrative and academic support for the major languages of the Global South, and at least financial aid and academic backing for the minority languages of the Global North. Such existing frameworks of support need to be taken into account when setting up collaborative efforts.

5 A DIVERSITY-AWARE LEXICAL MODEL

As a case study on the *value-sensitive design* [20] of language technology for the preservation of linguistic and cultural diversity, we present the *Universal Knowledge Core* (UKC), a large-scale multilingual lexical database, described from a technical perspective in [7, 24]. The UKC adopts a diversity-aware lexical concept space that is able to represent concepts that are culturally or linguistically specific to languages and communities, avoiding the type of bias illustrated in Figure 1.

Figure 3 shows the high-level lexical model of the UKC, using the example lexical field around the concept of rice, known to be culturally significant and diverse in several parts of the world. Horizontally, the model is divided into two layers: the *lexico-semantic layer* represents lexical meaning through concepts and their relationships (broader–narrower, part-of, etc.). The *lexical layer* represents the actual lexicalisations of these concepts, as well as lexical relations between them. Vertically, the model is divided into an *interlingua* (in yellow) that models unity, i.e. shared phenomena across languages, as well as one lexicon per language (in blue) that models diversity.

This bidimensional structure delineates four types of lexical knowledge: (1) lexico-semantic unity; (2) lexico-semantic diversity; (3) lexical unity; and (4) lexical diversity. (1) As shown in Figure 3, *lexico-semantic unity* asserts that the French *riz* and the Italian *riso* are equivalent, as both are connected to CONCEPT A of *rice*. Likewise, the Swahili *mchele* and the Japanese 米 are connected to the interlingual CONCEPT C of *uncooked rice*, which is asserted by the network to be a narrower term than *rice*, helping both humans and machines in its interpretation. (2) Lexico-semantic diversity provides evidence on untranslatability (e.g. no word exists for *rice* in Swahili) via explicitly representing *interlingual gaps*. These help MT systems identify difficult-to-translate phrases and handle them appropriately. Language-specific *local concepts*, as another form of diversity, are not merged into the interlingua, such as the Japanese *raw brown rice* in Figure 3. Through them, the UKC acknowledges the difficulty of integrating all culturally specific concepts from all societies into a single, coherent, global view. Even so, these local hierarchies remain connected to the interlingua through their root concepts, and can be exploited by applications destined to local communities. (3) *Rice*, *riz*, and *riso* do not only mean the same thing, they are also similar as word forms and are from a common etymological origin. The UKC models such *lexical unity* through *cognate* relationships. Cognates are a key tool in linguistic typology and lexicostatistics for the study of the similarity of lexicons [26]. They are also used in cross-lingual NLP applications, such as for building bilingual word embeddings [1]. (4) Finally, morphological and semantic information that relate to the form of the word, such as derivation or antonymy relations between words, are modelled as *lexical diversity*. Japanese and Italian both lexicalise *rice in the husk* (also called *paddy* in English); Italian, however, expresses this concept through derivation, via the augmentative *riso* → *risone*.

As of early 2024, the UKC contains about 1.9 million words from over 2,100 languages.[7] It expresses lexico-semantic and lexical unity through 111k interlingual concepts, 109k lexico-semantic and 3.3M cognate relations. Lexico-semantic and lexical diversity, in turn, are expressed via 39k interlingual gaps from 744 languages and over 230k language-specific relations including derivation,

[5] Retrieved in February 2023 from https://meta.wikimedia.org/wiki/List_of_Wikipedias.
[6] Kornai [37] has quantitatively proven a very strong correlation between simple measures such as Wikipedia presence and the general digital vitality of a language. For this reason, we consider the number of Wikipedia pages as a decent estimate for the overall digital content available in a language.
[7] The UKC does not contain named entities as it is not intended to be an encyclopedic resource nor to tackle issues typical of such resources.

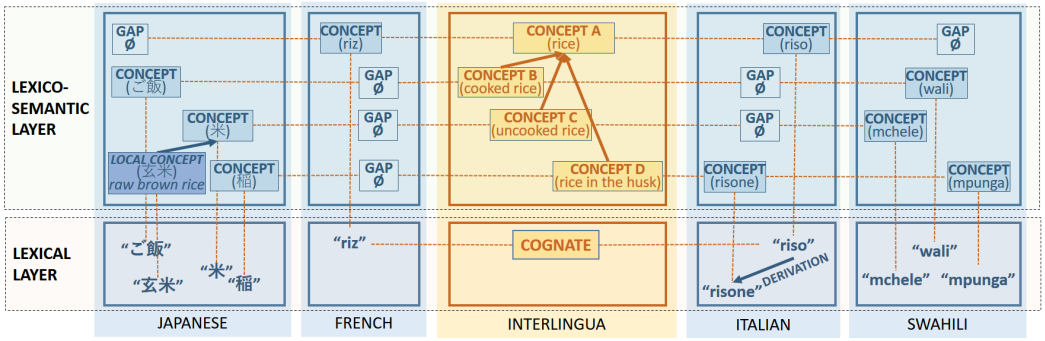


Figure 3: The cross-lingual mapping model of the UKC lexical database

antonymy, and metonymy. When computed on identical data (obtained from [24]) and in an identical manner as for the other four multilingual lexical databases in Section 3, the language modelling bias of the UKC lexical model is obtained to be zero. This is expected as the model was designed to solve interlingual mapping limitations present in other databases.

Content-wise, however, the UKC is not free of bias. First of all, it provides an uneven coverage of languages: only 7% of the lexicons have more than 1000 words, while major European languages have lexicons of more than 50 thousand words. This situation reflects the general state of multilingual language resources (many of which the UKC also incorporates). Secondly, the graph of interlingual concepts inside the UKC, having been bootstrapped from the concept hierarchy of the English Princeton WordNet, represents an Anglo-Saxon point of view on the conceptualisation of the world. Consequently, our recent *LiveLanguage* projects, following the methodology described in the next section, are focused on reducing these two kinds of bias within the UKC: by increasing the coverage of small lexicons, and by collecting and integrating evidence of language-specific words and untranslatability. Doing so, we are gradually shifting the resource from its Western perspective towards the needs of linguistic minorities and cultures of the Global South.

Past LiveLanguage projects on increasing UKC coverage involved lexicon extension on Scottish Gaelic [8], Mongolian [3], and the languages of India [17]. Ongoing projects are extending the lexicons of languages of Indonesia (Indonesian, Banjar, Javanese), South Africa (Setswana), as well Arabic [19].[8]

Our efforts on untranslatability focus on lexical domains known to be diverse across cultures. [35, 36] describe the redesign of the kinship domain of the UKC interlingua to allow it to represent over 2,00 kin terms and over 38,000 interlingual gaps in over 700 languages, relying both on ethnographic data [42] and our own field work. This effort led to extending the size of the kinship domain from a few dozen concepts mostly relevant to English to over 250 linguistically diverse concepts. Current work has moved to the domains of food and colour terms: while the diversity of food-related vocabulary is obvious, the different ways languages divide the colour spectrum via *basic colour terms* has also been widely discussed in linguistics [34].

[8]See the complete list of projects on http://ukc.datascientia.eu/projects

6 A DIVERSITY-AWARE DEVELOPMENT METHODOLOGY

The aims of the *LiveLanguage* initiative are to provide technical and methodological support for collaborative efforts on diversity-aware resource and tool development, and to disseminate the results of such efforts. According to the ethics policy of LiveLanguage, local communities set their goals and keep the intellectual property of all results. These principles constitute an ethical minimum to avoid an exploitative relationship, but are also justified by efficiency: they ensure that the project is useful and relevant, and that it motivates the local community in engaging with the project. A local institutional framework greatly simplifies the collaboration process: the local institution can act as the IP owner and has the necessary structure and network of people to organise the local effort.

For the co-creation of lexical resources, LiveLanguage adopts the six methodological steps shown in Figure 4.

(1) Project specification is led by the local institution, with support from LiveLanguage providing consultancy and know-how from past projects. It first determines (a) the long-term *goals and motivations* of the local community: language teaching, the development of AI-based language tech, basic language tools for smartphones, the preservation of cultural heritage, etc. Crucial design choices stem from these long-term goals: (b) the trade and local languages or dialects, linguistic phenomena, and semantic domains covered. For instance, a project motivated by teaching an endangered language to children will concentrate on building resources on the core vocabulary and basic grammar, while a project on the preservation of local culture may focus on the language of a specific domain such as food.

The choice of languages and domains determines, in turn, (c) the types and level of relevant linguistic and domain expertise needed to complete the project. The actors fulfilling key roles (e.g. language expert, domain expert, data collector, data validator) are recruited, possibly involving crowd or algorithmic workers. In line with the multipolar model of trade and local languages, a hierarchical organisation of the project is also determined: a main local institution is charged with project coordination, while its local partners (individuals or institutions) are in charge of efforts with respect to each local language. Finally, (d) the tools to be used by the actors and that are necessary to fulfil the goals are specified and their availability assessed. The specification process that starts with goals and then moves to languages and domains, actors, and finally to tools may

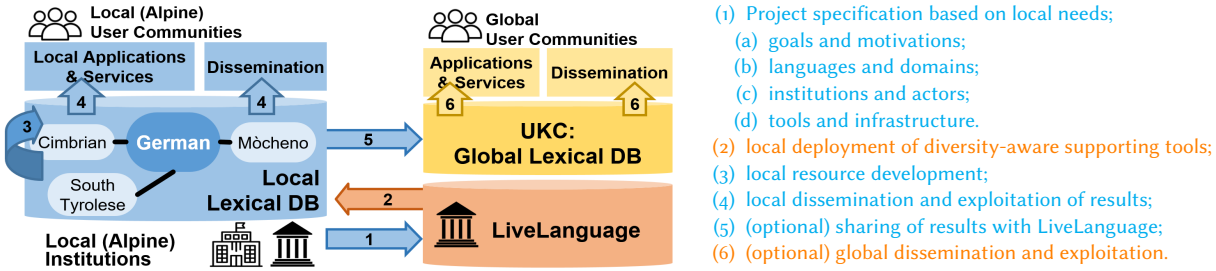


Figure 4: Collaborative language development methodology in the framework of the LiveLanguage initiative

not be linear: the non-availability of relevant expertise, workforce, or tools may lead to a reformulation of the goals to suit reality.

Let us take the example of a project on the languages and dialects within the Italian Alps, the goal of which is to expand existing lexicons for the purpose of teaching Alpine minority languages to local children. In this context, the trade languages are German and Italian, the local languages and dialects can be the (Germanic) Mòcheno, South Tyrolean, and Cimbrian, or the (Romance) Ladin or Friulian, and the domain is general language and the base vocabulary. The actors are mostly human experts due to the scarcity of existing language resources on which to base algorithms, and due to the small speaker communities: language teachers, university students, and civil enthusiasts. A hub university is overseeing the process while local language centres are in charge of recruiting and managing local actors.

(2) Deployment of diversity-aware supporting tools. Upon request, LiveLanguage can provide software and hardware tools and infrastructure to the local institution according to the needs determined in the project specification step, along with training and consultancy for the development process. In Figure 4, for an Alpine project, a *Local Lexical DB* is generated with the German lexicon as a trade language, and the existing (preliminary and incomplete) lexicons of three local languages, all automatically downloaded from the UKC via the LiveLanguage data catalogue. At the current stage, LiveLanguage provides the following software support:

- download of mono- or multilingual lexicons in a standard format from the LiveLanguage data catalogue;[9]
- a simple-to-use, open source lexical DB management system, automatically preloaded with existing UKC lexical data on the hub and satellite languages, mapped together as illustrated in Figure 3;
- browsing and visualisation tools for the multilingual lexicons, such as local versions of the UKC website[10];
- tools for the editing of lexicons.

(3-4) Local development, dissemination, and exploitation. The local institution(s) manage the process of resource development. They involve local collaborators according to project needs and, if necessary, may ask for consultancy (typically free of charge) from LiveLanguage. As they get to keep intellectual property rights over the resources produced, they have freedom to define the IP policies that govern the use of results, as well as to disseminate or exploit them through local applications or services. LiveLanguage provides tools both for development and dissemination of results.

(5-6) Global integration and dissemination. Local institutions are encouraged to share project results with LiveLanguage in their own interest: LiveLanguage offers the added value of mapping local lexicons, by means of the Universal Knowledge Core as a global lexical database, to all other languages. Local efforts thus provide the element of *diversity* into the lexicon, and integration with the UKC provides interlingual *unity*. The appearance of local lexicons in the UKC and the LiveLanguage data catalogue also provides an additional dissemination opportunity for the effort.

7 CONCLUSIONS AND FUTURE WORK

While there are notable efforts to address the digital language divide, and indeed the digital representation of previously underserved languages is increasing, these efforts tend to sacrifice linguistic and cultural specificities that have no equivalent in the world's most dominant languages. Having taken a stance for the development of technologies and methodologies that preserve such diversity, we have introduced *language modelling bias* as a quantitative measure of a technology's (in)ability to support languages in an equal manner. While recognising that technology with a completely bias-free representation of the world's languages is unattainable, our case study demonstrates how a language resource development project can take concrete steps towards avoiding discriminatory biases. The UKC and the LiveLanguage initiative are long-term projects that address both linguistic diversity and language modelling bias on the technological, methodological, ethical, practical, and social levels. The UKC was released to the global public in 2021 and has since been expanded with large amounts of lexical data on diversity. The collection of such data is a never-ending challenge, and we are initiating more projects, as well as offering new tools and services in the near future.

The collaborative process described in the previous section implicitly assumes the existence of a central organisation taking care of the long-term maintenance and sustainability of key LiveLanguage components: the UKC database and website, the LiveLanguage data catalogue, as well as the tools and services. While currently the University of Trento is playing this role, our short-term plans involve the creation of the DataScientia Foundation.[11] as a not-for-profit coordinating body where diverse stakeholders share decisional and operational power, incuding an international advisory board featuring experts from various language communities.

[9] https://datascientiafoundation.github.io/LiveLanguage/.
[10] See, for example, http://indo.ukc.datascientia.eu.
[11] http://datascientia.eu

ACKNOWLEDGMENTS

We thank the referees for their comments that helped improve this paper considerably. We also thank the Data Science Centre at University of Amsterdam that partly funded this research. Most of all, we express our gratitude to Adriano Clayton da Silva, the members of the Núcleo de Estudos de Linguagens da Amazonia (Amazon Language Studies Center), Khuyagbaatar Batsuren (Mongolian), Temuulen Khishigsuren (kinship), Hadi Khalilia (Arabic), Gina Tebatso Moape, and Sunday Olusegun Ojo (Setswana) who have provided us with examples of linguistic diversity which they consider particularly meaningful in terms of the cultural richness that finds expression in them.

REFERENCES

[1] Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2289–2294. https://doi.org/10.18653/v1/D16-1250

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (Eds.). Association for Computational Linguistics, Ann Arbor, Michigan, 65–72. https://aclanthology.org/W05-0909

[3] Khuyagbaatar Batsuren, Amarsanaa Ganbold, Altangerel Chagnaa, and Fausto Giunchiglia. 2019. Building the mongolian wordnet. In *Proceedings of the 10th global WordNet conference*. 238–244.

[4] Khuyagbaatar Batsuren, Ekaterina Vylomova, Verna Dankers, Tsetsuukhei Delgerbaatar, Omri Uzan, Yuval Pinter, and Gábor Bella. 2024. Evaluating Subword Tokenization: Alien Subword Composition and OOV Generalization Challenge. arXiv:2404.13292 [cs.CL]

[5] David Beer. 2017. The Social Power of Algorithms. *Information, Communication & Society* 20, 1 (2017).

[6] Gábor Bella, Khuyagbaatar Batsuren, Temuulen Khishigsuren, and Fausto Giunchiglia. 2022. Linguistic Diversity and Bias in Online Dictionaries. *University of Bayreuth African Studies Online* (2022), 173.

[7] Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Cheema, and Fausto Giunchiglia. 2022. Language Diversity: Visible to Humans, Exploitable by Machines. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 156–165.

[8] Gábor Bella, Fiona McNeill, Rody Gorman, Caoimhín Ó Donnaíle, Kirsty MacDonald, Yamini Chandrashekar, Abed Alhakim Freihat, and Fausto Giunchiglia. 2020. A major wordnet for a minority language: Scottish Gaelic. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 2812–2818.

[9] Emily M Bender. 2011. On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology* 6 (2011).

[10] Pushpak Bhattacharyya. 2010. IndoWordNet. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

[11] Steven Bird. 2020. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3504–3519. https://doi.org/10.18653/v1/2020.coling-main.313

[12] Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 7817–7829. https://doi.org/10.18653/v1/2022.acl-long.539

[13] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[14] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[15] Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1352–1362.

[16] Francis Bond, Luis Morgado da Costa, Michael Wayne Goodman, John Philip McCrae, and Ahti Lohk. 2020. Some Issues with Building a Multilingual Wordnet. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 3189–3197. https://aclanthology.org/2020.lrec-1.390

[17] Nandu Chandran Nair, Rajendran S. Velayuthan, Yamini Chandrashekar, Gábor Bella, and Fausto Giunchiglia. 2022. IndoUKC: A Concept-Centered Indian Multilingual Lexical Resource. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 2833–2840. https://aclanthology.org/2022.lrec-1.303

[18] Nicole Chi, Emma Lurie, and Deirdre K. Mulligan. 2021. Reconfiguring Diversity and Inclusion for AI Ethics. *https://arxiv.org/pdf/2105.02407* (2021). https://doi.org/10.1145/11952.107

[19] Abed Alhakim Freihat, Hadi Khalilia, Gábor Bella, and Fausto Giunchiglia. 2024. Advancing the Arabic WordNet: Elevating Content Quality. In *6th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT), LREC-COLING*.

[20] Batya Friedman, Peter Kahn, and Alan Borning. 2002. Value sensitive design: Theory and methods. *University of Washington technical report* 2, 8 (2002).

[21] Batya Friedman and Helen Nissenbaum. 1996. Bias in computer systems. *ACM Transactions on Information Systems* 14, 3 (Jul 1996), 330–347. https://doi.org/10.1145/230538.230561

[22] Lisa Gitelman. 2013. *"Raw Data" Is an Oxymoron*. MIT Press, Cambridge, MA, USA.

[23] Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and Exploiting Language Diversity.. In *IJCAI*. 4009–4017.

[24] Fausto Giunchiglia, Gábor Bella, Nandu C Nair, Yang Chi, and Hao Xu. 2023. Representing interlingual meaning in lexical databases. *Artificial Intelligence Review* (2023), 1–17.

[25] Joseph H Greenberg. 1956. The measurement of linguistic diversity. *Language* 32, 1 (1956), 109–115.

[26] Sarah C Gudschinsky. 1956. The ABC's of lexicostatistics (glottochronology). *Word* 12, 2 (1956), 175–210.

[27] Donna Haraway. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14, 3 (1988), 575. https://doi.org/10.2307/3178066

[28] Sandra Harding. 1995. "Strong Objectivity": A Response to the New Objectivity Question. *Synthese* 104, 3 (1995), 331–349.

[29] Paula Helm, Gábor Bella, Gertraud Koch, and Fausto Giunchiglia. 2024. Diversity and Language Technology: How Language Modeling Bias Causes Epistemic Injustice. *Ethics & Information Technology* (2024). https://doi.org/10.1007/s10676-023-09742-6

[30] Paula Helm, Loizos Michael, and Laura Schelenz. 2022. Diversity by Design? Balancing the Inclusion and Protection of Users in an Online Social Platform. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*. Association for Computing Machinery, New York, NY, USA, 324–334. https://doi.org/10.1145/3514094.3534149

[31] Ester Hlavnova and Sebastian Ruder. 2023. Empowering Cross-lingual Behavioral Testing of NLP Models with Typological Features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 7181–7198. https://doi.org/10.18653/v1/2023.acl-long.396

[32] Lilly Irani, Janet Vertesi, Paul Dourish, Kavita Philip, and Rebecca E. Grinter. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1311–1320. https://doi.org/10.1145/1753326.1753522

[33] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 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*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6282–6293. https://doi.org/10.18653/v1/2020.acl-main.560

[34] Paul Kay, Brent Berlin, Luisa Maffi, William R Merrifield, and Richard Cook. 2009. *The world color survey*. CSLI Publications Stanford, CA.

[35] Hadi Khalilia, Gábor Bella, Abed Alhakim Freihat, Shandy Darma, and Fausto Giunchiglia. 2023. Lexical diversity in kinship across languages and dialects. *Frontiers in Psychology* 14 (2023).

[36] Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using Linguistic Typology to Enrich Multilingual Lexicons: the Case of Lexical Gaps in Kinship. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. 2798–2807.

[37] András Kornai. 2013. Digital language death. *PloS one* 8, 10 (2013), e77056.

[38] Constantine Lignos, Nolan Holley, Chester Palen-Michel, and Jonne Sälevä. 2022. Toward More Meaningful Resources for Lower-resourced Languages. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, Dublin, Ireland, 523–532. https://doi.org/10.18653/v1/2022.findings-acl.44

[39] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment.

In *Socially Responsible Language Modelling Research*.
[40] Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What Kind of Language Is Hard to Language-Model?. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 4975–4989. https://doi.org/10.18653/v1/P19-1491
[41] George A Miller. 1998. *WordNet: An electronic lexical database*. MIT press.
[42] George Peter Murdock. 1967. Ethnographic atlas: a summary. *Ethnology* 6, 2 (1967), 109–236.
[43] Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. 216–225.
[44] Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2144–2160.
[45] Nanjala Nyabola. 2018. *Digital Democracy, Analogue Politics: How the Internet Era is Transforming Politics in Kenya*. Zed Books, London.
[46] Oxford Internet Study. 2015. The Digital Language Divide. http://labs.theguardian.com/digital-language-divide/.
[47] Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 3532–3542. https://doi.org/10.18653/v1/N19-1356
[48] Jan Rijkhoff, Dik Bakker, Kees Hengeveld, and Peter Kahrel. 1993. A method of language sampling. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"* 17, 1 (1993), 169–203.
[49] Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How Good is Your Tokenizer? On the Monolingual Performance of Multilingual Language Models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 3118–3135. https://doi.org/10.18653/v1/2021.acl-long.243
[50] Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Dublin, Ireland, 724–731. https://doi.org/10.18653/v1/2022.acl-short.82
[51] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672 [cs.CL]
[52] UNESCO. 2005. The Convention on the Protection and Promotion of the Diversity of Cultural Expressions. *https://en.unesco.org/creativity/convention* (2005).
[53] Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 2203–2213. https://doi.org/10.18653/v1/2021.eacl-main.188
[54] Steven Vertovec. 2012. "Diversity" and the Social Imaginary. *European Journal of Sociology* 53, 3 (2012), 287–312. https://doi.org/10.1017/S000397561200015X
[55] Jennifer C. White and Ryan Cotterell. 2021. Examining the Inductive Bias of Neural Language Models with Artificial Languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 454–463. https://doi.org/10.18653/v1/2021.acl-long.38
[56] Iris Marion Young. 1990. *Justice and the Politics of Difference*. Princeton University Press. Google-Books-ID: Q6keKguPrsAC.
[57] Michael J. Zimmerman and Ben Bradley. 2019. Intrinsic vs. Extrinsic Value. In *The Stanford Encyclopedia of Philosophy* (Spring 2019 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.