How Should We Model the Probability of a Language?

Rasul Dent[1], Pedro Ortiz Suarez[2], Thibault Clérice[1], Benoît Sagot[1]
[1]Inria, Paris, {firstname.lastname}@inria.fr
[2]Common Crawl Foundation, Paris, pedro@commoncrawl.org

Abstract

Of the over 7,000 languages spoken in the world, commercial language identification (LID) systems only reliably identify a few hundred in written form. Research-grade systems extend this coverage under certain circumstances, but for most languages coverage remains patchy or nonexistent. This position paper argues that this situation is largely self-imposed. In particular, it arises from a persistent framing of LID as decontextualized text classification, which obscures the central role of prior probability estimation and is reinforced by institutional incentives that favor global, fixed-prior models. We argue that improving coverage for tail languages requires rethinking LID as a routing problem and developing principled ways to incorporate environmental cues that make languages locally plausible.

1 Introduction

To use many Natural Language Processing (NLP) systems, we must first specify the language. The task of inferring this information is known as automatic language identification (LID). Following substantial progress in the 1990s, McNamee (2005) famously described LID for lengthy European-language documents as "a solved problem suitable for undergraduate instruction." Attention has since shifted to closely-related varieties (Aepli et al., 2023), rare languages (Caswell et al., 2020), short texts (Murthy and Kumar, 2006), and code-switching (Burchell et al., 2024).

Despite improvements on benchmarks, the practical reality for most languages has changed little in the past 20 years. New models are occasionally released for a region or family (e.g. Adebara et al., 2022), and some attempt to reach "the next 1000 languages" (e.g. Brown, 2013; Kargaran et al., 2023). Nonetheless, production-grade LID systems at companies like Meta and Google still focus on only a few hundred widely spoken languages.

This stagnation reflects two core issues. First, in research settings, LID is typically framed as a decontextualized inference task, in which systems aim to map directly text to a label from a global set. This encourages approaches that perform well on benchmarks but fail when applied to real data. Second, field-level incentive structures reward developing *novel* methods over revisiting basic tools.

We argue that recentering context will be crucial for expanding the *effective* coverage of LID. For "local languages" (see Bird, 2022) in particular, success should be defined relative to local contexts and resource constraints. This entails rethinking the values that shape research in LID.

2 The Received Framing of LID

At its core, LID is intended to route content to users who manage different languages. Due to differences between the modalities of text, speech, and sign, LID in each modality is generally approached separately, with text-based LID typically viewed as the easiest (Jauhiainen et al., 2024). For Rau (1974), the prototypical LID operator was a data entry clerk filing documents in unfamiliar languages for human patrons. With the explosion of born-digital content, the immediate recipient is now often a computer program.[1]

2.1 Standard Approaches

LID is typically approached as a supervised classification problem (Jauhiainen et al., 2024). This framing builds in two critical assumptions:

1. Labels represent a global hypothesis space.

2. Inference can be performed solely from text.

As in other classification tasks, there are two main modeling approaches. Per-class approaches

---

[1]However, Rau's vision still lives on in some institutions, like libraries.

such as Multinomial Naïve Bayes (e.g. Lui and Baldwin, 2012) fit the data to each class independently, and take the best fitting class as the label.[2] In contrast, discriminative models like fastText (Joulin et al., 2017; Grave et al., 2018) and CLD3 (Salcianu et al., 2023) learn decision boundaries between all classes simultaneously.

Hierarchical architectures also garnered interest. Many models, such as IndicLID (Madhani et al., 2023), are region-specific and depend on external software to ensure that they are used over the correct language set. Others, especially those based on fastText such as OpenLID (Burchell et al., 2023) and GlotLID (Kargaran et al., 2023), take a flat approach and model a wide variety of languages using one combined model. Yet others, such as LIMIT (Agarwal et al., 2023), try to learn hierarchical classification schemas directly from model errors.

2.2 Alternative Framings

The standard assumptions, and the modeling constraints that come with them, are not unreasonable in many common scenarios. In particular, they produce strong results on well-written monolingual documents in widely spoken languages (McNamee, 2005). However, their reliability deteriorates in several scenarios. Caswell et al. (2020) and Kreutzer et al. (2022) showed that the models that perform the best on common benchmarks often struggle when applied to noisy web data. Similarly, when working with closely-related varieties, or very short texts, it is often impossible to select just one correct label on the basis of input alone. The recent Shared Task on Improving Language Identification for Web Text at the 1st Workshop on Multilingual Data Quality Signals confirmed these issues remain relevant (Suarez et al., 2026).

In response to such issues, alternate framings have been proposed. For example, Baimukan et al. (2022) show that hierarchical labels are important for fine-grained dialect classification. Bernier-Colborne et al. (2023) and Keleg and Magdy (2023) extend this insight, reconceptualizing the task of dialect identification as one of multi-label classification. From a different angle, Dent et al. (2025) contend that building web corpora for rare languages is more of a mining task than strict classification.

Together, these reframings touch on a much more general issue, but do not completely resolve

it. Namely, general-purpose LID models need a label set fixed enough to allow training and flexible enough to handle great variation in inference-time granularity. In Section 3, we argue that the diversity of inference-time conditions has often been overlooked, leading to structural limitations with significant practical and scientific costs.

3 Probability Problems

Both per-class and discriminative approaches to LID ultimately estimate a conditional probability distribution: given input features X, estimate the probability P of each language ℓ, and choose the most likely label. Bayesian framing reveal that $P(\ell \mid X)$ actually depends on two terms, the probability of X given ℓ, and the prior probability of ℓ itself. The decision rule can be written in the familiar argmax form:

$$\hat{\ell} = \arg\max_{\ell} \ P(X \mid \ell)\, P(\ell), \qquad (1)$$

or equivalently in log-space:

$$\hat{\ell} = \arg\max_{\ell} \left[\log P(X \mid \ell) + \log P(\ell) \right]. \qquad (2)$$

This decomposition highlights a highly non-trivial modeling problem. Namely, what does $P(\ell)$ actually represent?

3.1 Global Frequency?

In the classic Bayesian formulation, the answer is simple enough. $P(\ell)$ is the number of texts in a given language over the number of examples in the entire corpus. In principle, this should correspond to a real difference in the frequency of languages, rather than a sampling artifact. However, under naïve frequency-based estimates, we immediately face two related problems when dealing with massive class imbalances.

1. Rare classes become nearly undetectable.

2. Ignoring rare classes has minimal effect on macro-performance.

To see this concretely, suppose we compare English ($\ell = $ en) with a rare language ($\ell = r$). Let the global priors be:

$$P(\text{en}) = 0.40, \qquad P(r) = 10^{-6}.$$

Assume for illustration that the likelihoods of a given text under the two languages are of comparable magnitude—for example:

---

[2]Traditionally, this is called generative modeling, but as an anonymous reviewer noted, the term has become ambiguous.

$$P(X \mid \text{en}) = 10^{-4}, \qquad P(X \mid r) = 10^{-2}.$$

Even though the rare-language model assigns a *hundred times higher* likelihood to the input text, the posterior still overwhelmingly favors English:

$$P(X \mid \text{en})P(\text{en}) = 10^{-4} \cdot 0.40 = 4 \times 10^{-5},$$

$$P(X \mid r)P(r) = 10^{-2} \cdot 10^{-6} = 10^{-8}.$$

Thus,

$$4 \times 10^{-5} \gg 10^{-8},$$

and the argmax will select English. The only way the likelihood evidence can compensate for a prior that is effectively zero is if the likelihood for more common languages is *also* effectively zero (likely due to a script mismatch). Because language r occurs less than a thousandth of the time, getting it completely wrong will barely show up as a rounding error in overall accuracy.

3.1.1 Attenuated Frequency?

Due to the extreme distribution of true frequencies, one common mitigation strategy is to upsample rare classes and/or downsample common ones (Burchell et al., 2023). This can help prevent a low $P(\ell)$ from instantly overpowering the likelihood evidence. However, even with this, there is a double bind. On the one hand, the prior still favors classifying data without obvious hints under a dominant class. On the other hand, the model can rely on a few shallow features and thus lack robustness when such features appear as noise in real data. These two issues have led recent studies to emphasize False Positive Rate (FPR) as a key metric.

3.1.2 False Positives at Scale

To contextualize this issue, we revisit the example of Caswell et al. (2020), who note that a LID classifier that upsamples r enough to achieve 99% precision and 99% recall on a balanced data, could be expected to recover 9.9K out of 10K true positives. However, if this is applied at web scale over 100B documents with a 0.01% FPR, the resulting 10M false positives would render the resulting corpus unusable. Thus, adjusting training data ratios and carefully choosing metrics can reduce the damage, but they do not solve the underlying issue.

Naturally, one is tempted to ask where these false positives come from. While the obvious answers of high-resource and out-of-model cousins do indeed play a role, Caswell et al. (2020) also document various forms of noise that are difficult to *thoroughly* account for at training time. While a full review is beyond the scope of this article, important classes include emoji, misrendered PDFs, non-Unicode fonts, chance n-gram overlaps between unrelated languages, 'A N T S P E A K', and mixed-script boilerplate. In response to this, they, as well as later filtered crawls like Fineweb-2 (Penedo et al., 2025), emphasize *post-filtering* for document-level consistency and token-precision in minority languages. Once again, this mitigates the fallout, but does not solve the issue.

3.2 Local Priors?

Once we appreciate the magnitude of the problem of false positives at scale, the central question is no longer how well languages can be discriminated in the abstract, but rather how to avoid suggesting labels that are not even plausible candidates in a given context. We must first remember that, even for widely spoken languages, global frequency is often irrelevant. For instance, English, Mandarin, and Russian each account for a substantial share of web text, but each is typically written in a distinct script. Observing the script instantly render demographic dominance of the others nearly meaningless, and Bayesian models like CLD2 often incorporate preliminary script classification for this very reason.

In theory, similar effects should arise when incorporating metadata such as geography, social networks, or user feedback. These contextual signals are useful in general, but crucial for assigning usable prior probabilities to local languages, as well as managing the sheer size of the label inventory. In practice, however, converting context into usable prior probabilities remains challenging. Although some recent work has explored the utility of geodata (Dunn and Edwards-Brown, 2024), current benchmarks make it difficult to compare such context-aware models. The challenge is not simply that contextual information is missing; rather, the logistical requirements are fundamentally incompatible with how multilingual datasets are generally collected, labeled, and shared today.

3.2.1 Where Do Languages Live?

Firstly, there is the question of whether tail languages are actually used enough that misclassification is a practical, rather than theoretical, issue. Recently, many languages that have traditionally been considered "unwritten", ranging from European varieties like Luxembourgish (Wagner, 2013) to the Zapotec languages of Mexico (Lillehaugen, 2019), *are* increasingly written (and spoken) on social media platforms and messaging services. Over the last few years, both and privacy regulation and contractual restrictions have curtailed research access, but there is reason to believe this trend continues.

3.2.2 Dataset Difficulties

Given the limited availability of user data, public LID datasets turn to sources that are easier to work with, like Wikipedia, the Bible, or newswire.[3] Here, providing metadata about sources, by saying for example that the passage comes from a Saint Lucian Bible, would often solve the task, so the source is not given. Even the full, document-length text would still be to easy, so the tendency is to simulate harder scenarios by decomposing data from these sources into short, freestanding snippets, as in FLORES-200 (NLLB_Team et al., 2022) and GlotLID-C (Kargaran et al., 2023).

These practices can lead to a non-trivial circularity problem. Creating reliable labels for languages beyond the top 100–200 demands annotators with broad linguistic coverage, including familiarity with lesser-known scripts and regional varieties, as exemplified by Kreutzer et al. (2022). Even when such expertise can be assembled, annotators typically lack access to contextual metadata for the reasons we have just discussed. In practice, this is mitigated by beginning with pre-compiled corpora, which cover precisely the kinds of data for which LID already works.

3.2.3 Model Expectations

Older per-class models like CLD2 allowed users to provide hints like the expected language(s), URL sources, top-level domain name, original encoding, and embedded language tags. In contrast, newer discriminative classifiers like fastText and CLD3 that are deeply embedded in industrial and academic preprocessing pipelines cannot easily take local context into account. While it is possible to restrict the output space of such models to a few

---

[3] Although even these are increasingly the subject of copyright challenges

candidate languages, this involves an opaque manipulation of the softmax layer that can lead to unstable behavior.

3.2.4 Towards Context-Aware LID

In light of these operational difficulties, it is entirely understandable that public benchmarks do not focus on the relationship between the idealized task and real-world deployments. However, this encourages the development of models optimized for an version of LID that simultaneously overestimates the difficulty posed by rare languages in their local contexts and underestimates the impact of noise at scale. This makes it difficult to justify context-aware methods that target certain kinds of confusion that mainly surface when raw data is treated across diverse global sources. In Section 5, we will argue that this technical misalignment is reinforced by powerful institutional incentives, which further entrench the decontextualized framing of LID. However, we first present two case studies in Section 4 which help illustrate the problem.

4 Case Studies

As a brief demonstration, we briefly consider two contact languages where the global prior probability is effectively zero, resulting in exclusion from all SOTA models. The first is Louisiana Creole (LC), a critically endangered language used online in the context of language revitalization. The second is the now-extinct Lingua Franca, long suspected to have influenced contact languages around the world.

4.1 Language Revitalization

Following over a century of language shift to English, LC is spoken today by a few thousand people. However, there is a revitalization movement with a strong online presence on several social media platforms, such as Facebook and Discord. In revitalization spaces, the main languages are LC and English. Some groups tolerate discussion of potentially relevant languages (Louisiana) French, Spanish, and other French-based Creoles, while others, like the Louisiana Creole Virtual Classroom prefer a strictly bilingual learning environment (Mayeux, 2024). In any case, $P(LC)$ is much higher in dedicated groups than its global rarity would suggest.

However, as the LID label set for existing models is defined at training time, the modeling assumption is that $P(LC) = 0$. When downstream services like auto-translate are activated, the language

Figure 1: Interfaces allow override, but not for LC. (Google Translate)

is frequently misrecognized as French, which predictably leads to misunderstandings. Some platforms now have options to rate translations, guide the machine translation to a slightly closer target (i.e. Haitian Creole) or disable auto-translation. However, at the time of writing, there is no user-facing option to identify a sentence as LC, reflecting the absence of LC from the underlying LID label set. As a result, users can mitigate downstream effects but cannot directly correct the initial language identification.

To explore this issue, we present a sample from Google Translate in Figure 1 that is functionally similar to social media, but deliberately not taken from a real group due to privacy considerations. The input text is *Mo sòr fé ègzemp-çila pou montré kofè li marsh pa*[4]. The language is misidentified as Bambara and incorrectly translated as 'I'm sure I did example to show coffee li marsh pa'.

What is of interest here is not the mistranslation, which is expected since the language is not supported, but rather the LID interface. Previous work by (Yang et al., 2025a) has taken the absence of Native American languages as labels in Google Translate (and similar industrial models) as an indication that LID is a bottleneck. Yet, there *is* an option for manual override. Crucially, this only extends to fully supported languages like Haitian Creole and French. Moreover, we are able to create a shortlist of languages that have recently been used, and are thus presumed to appear again. Although the shortlist does not seem to bias the LID classifier, this interface showcases that for common languages, manual override is accepted and does not require a 7,000 language menu every time. Once we observe this capacity for manual selection, it becomes clearer that the question is not necessarily whether languages can be recognized, but rather whether recognition can be decoupled from full support. We return to this issue in Sections 5.2 and 6.3.

---

[4]'I just made this example to show why it does not work'

4.2 Retracing Language Contact

Whinnom (1956) once argued that the Mediterannean Lingua Franca (LF) may have been an inspiration for contact varieties in Africa and Asia. Although the general idea has been largely rejected since the 1980s, Nolan (2023) claims LF may have influenced the 19th-20th century French military register known as Français Tirailleur (FT). However, it has been extremely difficult to validate even this limited claim two main reasons. Firstly, LF was a rarely-written low status contact language based on — yet distinct from — prestige Romance languages like Italian and Spanish. Secondly, it was mostly used in North Africa and the Levant, where local writing was typically in the Arabic script and furthermore has not necessarily been digitized to the same extent.

Overall, current LID systems are not optimized for recovering historical contact varieties as heterogeneous as LF, and existing models do not attempt to. Even if they did, the small extant corpus and high lexical overlap with much better documented Romance varieties as well as Romanized Arabic would likely lead to confusion with these more frequent categories.

However, we have found reports which show that the French military served as a powerful vector for the global transmission of LF, also known as Sabir, following the French occupation of Algiers. In particular, the *Relation historique de la campagne du Sénégal (1861) : extrait du journal d'un capitaine de tirailleurs algériens*[5] (Dubois, 1863) includes several passages of dialogue in the contact variety during an expedition in Senegal around the time FT is thought to have developed. For example, *Toi courageux kif r'azaloun, toi parlar kif barougue, toi tornar M'Boro, toi fasir meillour Lackmar.*[6] In this excerpt alone, the pronouns *toi* 'you' is French, the names of the animals are Arabic, the verbs are Mediterranean Romance, and M'Boro is the name of a town in Senegal.

In a decontextualized setting, this combination would look like a code-switching nightmare. Yet, in the context of the 19th French *Armée d'Afrique* 'Army of [North] Africa' , the main task is to distinguish LF from French and to a much lesser extent, Arabic. This is a much more tractable sub-

---

[5]Historical Relation of the Senegal Campaign (1861) : Extract from the Journal of a Captain of the Tirailleurs Algériens

[6]You who are courageous like a gazelle and speak like a parrot, you go back to M'Boro, you will do better than Lackmar.

problem. In doing so, we identify a historically plausible mechanism of transmission in the form of Sabir-speaking Algerian troops transported to West Africa and placed in daily contact with newly conscripted Senegalese recruits. Although validating such historical connections is clearly feasible with lightweight NLP methods, such questions remain relatively marginal for reasons we explore in Section 5.

5   Why Is This Hard To Change?

Even if the limitations of decontextualized LID are clear, several distinct incentive structures across academia, industry, and public institutions prevent these problems from being addressed in practice.

5.1   Academic Incentives

Within academic NLP, it is often assumed that meaningful contributions should involve novel architectures, training strategies, or datasets. Demonstrating that a well-understood technique still works is typically viewed as incremental rather than interesting. As a result, researchers are structurally disincentivized from relying on simple or well-established methods. This pattern is reflected in persistent under-citation of earlier foundational work (Singh et al., 2023) and a broader tendency to treat re-engineering as innovation, had been amplified by the rapid growth of the field and the prestige hierarchy surrounding machine learning conferences.

More specifically, *Excitement* has been explicitly recognized as a key criterion for top-conference acceptance within the *ACL umbrella (Rogers et al., 2023). This is well suited to promoting major breakthroughs, but it also deprioritizes contributions that are merely *Sound*. Since LLMs are not yet a practical or cost-effective substitute for (shallow) general LID systems, there is an increasing gap between what the interests of the broader field and what actually works for this particular task.

To be clear, this dynamic does not strictly *prevent* work like van der Goot (2025) from being published in main conferences. However, it does tend to route LID papers toward Findings and especially specialized workshops. The issue is not the quality of these venues, but their position in the prestige hierarchy. Because such venues lack the visibility and career incentives associated with the main ACL-branded conferences, there is little incentive for early-career researchers to specialize in LID, as opposed to other, "more contemporary" problems.

5.2   Commercial Incentives

Equally, if not more important, is the reward structure of private industry. Here, the primary issue is not necessarily novelty or prestige. Rather, it is that existing LID solutions work well enough for commercially attractive languages, and improving LID for tail languages does not clearly lead to a viable revenue stream. Thus, even though there have been efforts such as Caswell et al. (2020), LID is not often treated as a pressing issue in its own right, but rather as a routing component for machine translation (e.g. NLLB_Team et al., 2022), as explored in Section 4.

Furthermore, accurately identifying a language implicitly commits a platform to a certain level of support for that language. Once a system signals that it can reliably recognize a language, users may reasonably expect corresponding improvements in areas such as content moderation, abuse detection, policy enforcement, and ranking. However, fully supporting even one additional language often requires coordinated work across multiple teams such as policy, annotation, trust and safety, and regional compliance. This organizational effort can exceed the technical effort by a significant margin.

From a business perspective, the combination of increasing operational and compliance costs with no clear additional revenue is particularly discouraging. This in turn casts a large shadow on the data asymmetry problem, since, as noted in Sections 3.2.1 and 4.1, industry platforms often contain precisely the high-quality, contextual data for rare languages necessary to build and evaluate robust LID systems.

5.3   (Absence of) Vertical Integration

In principle, progress in ASR, transcription, text processing, and LID should be mutually reinforcing. ASR systems require textual resources; text collection is far more efficient when ASR exists; and accurate LID is a prerequisite for building both. This feedback loop operates naturally for many large languages.

For local languages, however, the previously named incentives lead to structural deadlock. Each component waits for progress in the others, and only the largest actors have incentive to coordinate the pipeline. For example, some government and national security agencies manage to overcome

these issues and build vertically integrated systems, they cannot release them. It is true that some commercial endeavors like Omnilingual ASR team et al. (2025) are starting to pay more attention to these interdependencies, but they have yet extend them to textual LID. Thus, even though public radio archives and contemporary broadcasts in principle provide enough material to put many languages into medium-resource territory (Doumbouya et al., 2021; Havard et al., 2025), a lack of vertical integrations renders many of these resources effectively unusable for the moment.

6 What Can We Do?

In Section 3, we explored how using invariant priors poses special problems for languages that are globally rare but locally common. In Section 4 we saw that such languages present particular challenges across social media and archives that are not insurmountable at the local level. In Section 5, we explored some institutional obstacles to incorporating contextual cues in LID models. Now, we pivot from describing problems to outlining alternative solutions for these interwoven issues.

6.1 Two Coherent Paths Forward

In introductory computer science, we are explicitly taught to place correctness at the center of program design. A large part of correctness concerns preempting edge cases and testing them. In LID, we recognize that most languages are edge cases, and simply eliminate them the label space during training. This is often justified as a necessity caused by data and/or model limitations. But is doing so correct?

The answer depends on one's conception of LID. If one accepts, as we have observed, that LID typically functions to route content within a larger system, then success consists of reliably moving content from Source A to Target B. This framing naturally treats LID as an engineering problem where solutions are evaluated relative to downstream requirements. If, however, we wish to maintain that LID remains a scientific problem, then it is imperative that the determination of prior plausibility be treated as problem in its own right.

To make the role of prior specification concrete, we briefly sketch how local context can be incorporated into a standard per-class language identification model without introducing new architectures or training procedures. We define a *local frame of reference* as any mechanism that restricts the hypothesis space of candidate languages or modifies their prior probabilities before observing the text itself, based on external signals such as document provenance, interface-level language settings, discourse continuity, previous inferences, or explicit user hints. Within such a frame, language identification consists of selecting the most plausible language relative to a locally valid prior distribution, rather than with respect to a single, globally fixed prior over all languages.

Concretely, let us (re)consider the Naive Bayes LID model represented in Section 3. Under a local frame of reference, the global prior, which implicitly encodes assumptions about worldwide language frequency, is replaced by a context-dependent distribution $P_{\text{local}}(\ell)$, yielding

$$\hat{\ell} = \arg \max_{\ell \in \mathcal{L}_{\text{local}}} \left[\log P(X \mid \ell) + \log P_{\text{local}}(\ell) \right].$$

In the simplest case, this local context can act as a hard gate: languages deemed implausible in the current setting are removed from the hypothesis space, while remaining languages receive a uniform local prior. More intricate formulations allow for soft gating, in which locally plausible languages are upweighted relative to a global prior, while others retain a small but non-zero probability mass. Importantly, both variants substitute local priors for global ones at inference time, and do not require retraining the underlying language models. For coarse metadata on most languages, including alternate names, LinguaMeta (Ritchie et al., 2024) provides a usable starting point. More localized data, such as placing Native Alaskan languages in Alaska specifically, would likely help limit false positives (cf. Yang et al., 2025b).

We recognize that discrimination between very similar varieties remains challenging, and that even creating likelihood models requires a certain amount of data. This approach is fully compatible with hierarchical and multi-label setups that leverage specific similarities and differences between related languages for fine-grained classification.

6.2 Evaluation and Benchmarks

One implication of the fork outlined above is that current evaluation practices conflate two distinct questions: whether a system can discriminate among languages given a fixed hypothesis space, and whether that hypothesis space is itself appropriate for a given use case. Under both paths, this

conflation obscures the source of error and limits meaningful comparison.

As mentioned in Section 6.1, if LID is treated primarily as an engineering problem, then evaluation should be aligned with deployment conditions. Rather than optimizing for global accuracy over a fixed label set, systems should be assessed relative to the routing decisions they are intended to support, including downstream error tolerance, latency constraints, and the availability of auxiliary signals. In this setting, comparisons across platforms or domains are not expected to be universal. Instead, success is defined locally, in terms of whether content is routed to an appropriate handler with acceptable cost.

If, by contrast, LID is to remain a scientific task, then evaluation must explicitly separate prior determination from text-based inference. Benchmarks that present decontextualized text snippets implicitly encode an often arbitrary prior of convenience over the label space based on the available data. A scientifically meaningful alternative would require benchmarks in which non-linguistic signals are exposed as inputs and evaluated in their own right. This would allow models to be compared on their ability to restrict the hypothesis space before observing the text, as well as on their performance once that restriction has been applied.

Crucially, such benchmarks need not rely on sensitive or proprietary user data. As demonstrated by deployed systems, coarse signals such as interface-level language selection, document provenance, discourse continuity, or deliberately abstracted metadata can already produce substantial reductions in uncertainty. Making these signals explicit would not simplify the task; rather, it would clarify which aspects of LID difficulty arise from linguistic ambiguity and which arise from prior misspecification.

Under either interpretation, the prevailing practice of evaluating LID systems solely on context-free text classification conflates engineering convenience with scientific validity. The result is not merely suboptimal performance for long-tail languages, but an evaluation procedure that is unable to distinguish between failures of modeling and failures of task definition.

6.3 Transparency and User Interaction

As observed in Section 5, institutions have various commitments that can make it difficult or unnecessary to support every language uniformly. For web platforms in particular, successful LID without corresponding downstream support can lead to complicated operational and organizational situations. However, in many spaces where tail languages are actually used, it can appear that LID is inherently difficult, when in fact performance could be meaningfully improved through local context injection.

It is worth reiterating that the issue is not simply reduced accuracy for these languages. Rather, they are systematically eliminated from consideration altogether, often through implicit prior assumptions. This occurs despite the fact that mechanisms such as manual override or constrained language selection (potentially combined with verification) are already accepted for widely supported languages. As such, even when such compensatory mechanisms are not implemented, making the exclusion of a language explicit is important. Setting $P(\ell) = 0$ *everywhere* is a strong modeling decision, and leaving it implicit can lead to confusion when tested on languages where identification is often trivial due to script, such as Cherokee (Eggleston et al., 2025).

7 Conclusion

Despite the perception that language identification is an easy task, the majority of the world's languages are not still not consistently identified in writing. Correctly identifying such languages is a key step for reducing digital disparities across speech communities. One of the root causes of this gap is the difficulty of defining the local probability of a language. To move forward, we must pay special attention to the social contexts these languages are used in, and reconsider institutional practices that center novelty and modeling convenience.

Limitations

We recognize that empirical demonstrations may have helped to make some of our arguments more concrete. However, due to the very nature of SOTA discriminative classifiers, directly exploring the effects of dynamic priors would have taken us beyond the scope of this paper.

Ethical Considerations

Whether or not one agrees with our position, it should be clear that there is a certain tension between accurate coverage of highly localized language on one hand and privacy considerations on the other. In this paper, we have argued that using

metadata like especially community and discourse structure is likely necessary to overcome technical obstacles to identifying languages that are not covered by existing large-scale systems. This does not imply that LID *must* be implemented even in communities that express explicit concerns about privacy and/or surveillance. However, in cases where such considerations are a relevant factor, they should be stated alongside other engineering constraints.

Acknowledgments

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. AfroLID: A Neural Language Identification Tool for African Languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial Evaluation Campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language Identification, Misidentification, and Translation using Hierarchical Models in 350+ Languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2022. Hierarchical Aggregation of Dialectal Data for Arabic Dialect Identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4586–4596, Marseille, France. European Language Resources Association.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and Variant Identification as a Multi-Label Classification Task: A Proposal Based on Near-Duplicate Analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Steven Bird. 2022. Local Languages, Third Spaces, and other High-Resource Scenarios: 60th Annual Meeting of the Association for Computational Linguistics, ACL 2022. *ACL 2022 - 60th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, pages 7817–7829. Publisher Copyright: © 2022 Association for Computational Linguistics.

Ralf D. Brown. 2013. Selecting and Weighting N-Grams to Identify 1100 Languages. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Ivan Habernal, and Václav Matoušek, editors, *Text, Speech, and Dialogue*, volume 8082, pages 475–483. Springer Berlin Heidelberg, Berlin, Heidelberg.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An Open Dataset and Model for Language Identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Robert Thompson, and Kenneth Heafield. 2024. Code-Switched Language Identification is Harder Than You Think. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 646–658, St. Julian's, Malta. Association for Computational Linguistics.

Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rasul Dent, Pedro Ortiz Suarez, Thibault Clérice, and Benoît Sagot. 2025. Identifying Rare Languages in Common Crawl Data is a Needles-in-a-Haystack Problem. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1460–1473, Suzhou, China. Association for Computational Linguistics.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14757–14765.

E. Dubois. 1863. *Relation historique de la campagne du Sénégal (1861) : extrait du journal d'un capitaine de tirailleurs algériens*. P. Arnavon, Blidah.

Jonathan Dunn and Lane Edwards-Brown. 2024. Geographically-Informed Language Identification. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7672–7682, Torino, Italia. ELRA and ICCL.

Liam Enzo Eggleston, Michael P Cacioli, Jatin Sarabu, Ivory Yang, and Kevin Zhu. 2025. Bridging the

Gap: Leveraging Cherokee to Improve Language Identification for Endangered Iroquoian Languages. In *Proceedings of the 1st Workshop on Advancing NLP for Low-Resource Languag*, Varna, Bulgaria.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

William N. Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Self-Supervised Models of Speech Processing for Haitian Creole. *BABEL*, 1091(547):544.

Tommi Jauhiainen, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2024. *Automatic Language Identification in Texts*. Synthesis Lectures on Human Language Technologies. Springer International Publishing.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schuetze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6155–6218, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic Dialect Identification under Scrutiny: Limitations of Single-label Classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Brook Danielle Lillehaugen. 2019. Tweeting in Zapotec: Social Media as a Tool for Language Activists. In Jennifer Carolina Gómez Menjívar and Gloria E. Chacón, editors, *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*, pages 202–226. University of Arizona Press, Tuscon.

Marco Lui and Timothy Baldwin. 2012. Langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Yash Madhani, Mitesh M. Khapra, and Anoop Kunchukuttan. 2023. Bhasa-Abhijnaanam: Native-script and romanized Language Identification for 22 Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 816–826, Toronto, Canada. Association for Computational Linguistics.

Oliver Mayeux. 2024. Language revitalization on social media: Ten years in the Louisiana Creole Virtual Classroom. *Language Documentation and Description*, 23(1).

Paul McNamee. 2005. Language identification: A solved problem suitable for undergraduate instruction. *Journal of computing sciences in colleges*, 20(3):94–101.

Kavi Narayana Murthy and G. Bharadwaja Kumar. 2006. Language identification from small text samples. *Journal of Quantitative Linguistics*, 13(1):57–80.

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, and 20 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *Preprint*, arXiv:2207.04672.

Joanna Nolan. 2023. *Lingua Franca and Français Tirailleur: From Sea Jargon to Military Order?* Springer International Publishing, Cham.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. *Preprint*, arXiv:2511.09690.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. FineWeb2: One Pipeline to Scale Them All – Adapting Pre-Training Data Processing to Every Language.

Morton David Rau. 1974. *Language Identification by Statistical Analysis*. Ph.D. thesis, Monterey, California. Naval Postgraduate School.

Sandy Ritchie, Daan van Esch, Uche Okonkwo, Shikhar Vashishth, and Emily Drummond. 2024. LinguaMeta: Unified Metadata for Thousands of Languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10530–10538, Torino, Italia. ELRA and ICCL.

Anna Rogers, Marzena Karpinska, Jordan Boyd-Graber, and Naoaki Okazaki. 2023. Program Chairs' Report on Peer Review at ACL 2023. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages xl–lxxv, Toronto, Canada. Association for Computational Linguistics.

Alex Salcianu, Andy Golding, Anton Bakalov, Chris Alberti, Daniel Andor, David Weiss, Emily Pitler, Greg Coppola, Jason Riesa, Kuzman Ganchev, Michael Ringgaard, Nan Hua, Ryan McDonald, Slav Petrov, Stefan Istrate, and Terry Koo. 2023. Compact Language Detector v3 (CLD3). Google.

Janvijay Singh, Mukund Rungta, Diyi Yang, and Saif Mohammad. 2023. Forgotten Knowledge: Examining the Citational Amnesia in NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6192–6208, Toronto, Canada. Association for Computational Linguistics.

Pedro Ortiz Suarez, Laurie Burchell, Catherine Arnett, Rafael Mosquera-Gómez, Sara Hincapie-Monsalve, Thom Vaughan, Damian Stewart, Malte Ostendorff, Idris Abdulmumin, Vukosi Marivate, Shamsuddeen Hassan Muhammad, Atnafu Lambebo Tonja, Hend Al-Khalifa, Nadia Ghezaiel Hammouda, Verrah Otiende, Tack Hwa Wong, Jakhongir Saydaliev, Melika Nobakhtian, Muhammad Ravi Shulthan Habibi, and 78 others. 2026. CommonLID: Re-evaluating State-of-the-Art Language Identification Performance on Web Data. *Preprint*, arXiv:2601.18026.

Rob van der Goot. 2025. Identifying Open Challenges in Language Identification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18207–18227, Vienna, Austria. Association for Computational Linguistics.

Melanie Wagner. 2013. Luxembourgish on Facebook: Language Ideologies and Writing Strategies. *Social Media and Minority Languages: Convergence and the Creative Industries*, 152:87.

Keith Whinnom. 1956. *Spanish Language Contact in the Phillipines*. Hong Kong University Press, London.

Ivory Yang, Weicheng Ma, Chunhui Zhang, and Soroush Vosoughi. 2025a. Is It Navajo? Accurate Language Detection for Endangered Athabaskan Languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 277–284, Albuquerque, New Mexico. Association for Computational Linguistics.

Ivory Yang, Chunhui Zhang, Yuxin Wang, Zhongyu Ouyang, and Soroush Vosoughi. 2025b. Visibility as Survival: Generalizing NLP for Native Alaskan Language Identification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6965–6979, Vienna, Austria. Association for Computational Linguistics.