# Confusable Characters as Endangered Language Markers: The Case of North Caucasus Writing Systems

**Alexander Gutkin**° **Adrian Benton**† **Christo Kirov**† **Brian Roark**‡
Google Research, °London, UK; †New York; ‡Portland, OR
{agutkin,adbenton,ckirov,roark}@google.com

**Abstract**

The Abkhaz-Adyghe and Nakh-Daghestanian language families encompass 35 living languages that possess arguably the most complex modern Cyrillic orthographies due to their very sophisticated phonology. The relevant online data displays idiosyncratic patterns among which the use of confusable characters in input methods is the most prevalent. This work studies one such character—letter *palochka*—that is shared by most of the writing systems in question. We investigate whether patterns including variants of this character can act as markers of these languages in large-scale web-crawled data. We use GlotLID, a wide-coverage off-the-shelf language identification (LID) model, to label paragraph-level web text that contains a palochka confusable, and estimate the effect of confusable character normalization on the quality of GlotLID's predictions in 14 supported North Caucasian languages. According to GlotLID, the normalization significantly increases the recall (discovery of new language data) for some languages, while degrading it for others. However, manual evaluation reveals that overall, only 41% of ensuing wins and 46% of losses are accurate due to GlotLID prediction errors. We argue that, despite finding useful signals, higher precision LID approaches tailored to these long-tail languages are needed to improve the quality of mined data.

**Keywords:** endangered languages, language identification, homoglyphs, data mining, North Caucasus

## 1. Introduction

Large web-crawled text corpora in the Cyrillic script are becoming increasingly important as large natural language models are extended to cover lower-resource languages (Krsteski et al., 2025; Kuzhuget et al., 2024; Togmanov et al., 2025; Isbarov et al., 2025). For long tail languages, web-crawled sources offer a valuable alternative to traditional hand-curated datasets, which can be challenging to curate due to a scarcity of linguistic expertise and limited access to native speakers (Caswell et al., 2020; Bapna et al., 2022). Additionally, web data may play a particularly important role in the preservation and revitalization of endangered languages with Cyrillic writing systems, which often suffer from a limited online presence (Nikitina et al., 2019; Arkhangelskiy, 2019; Yankovskaya et al., 2023; Rueter et al., 2024). This last point has been emphasized even for the languages with federal republican status within the Russian Federation, such as Buryat, for which the relatively large number of native speakers does not correlate with the amount and quality of online language material (Khilkhanova, 2019; Zhanaev and Połeć, 2019). While sociolinguists often rely on social media data, general web-crawled corpora can provide a wider range of language data from more diverse sources, potentially including underrepresented Cyrillic orthographies.

To reach adequate quality levels of harvested data, aggressive filtering is typically applied, including setting thresholds on language identification (LID) confidence to retain text (Kudugunta et al., 2023; Kargaran et al., 2024). The quality and coverage of the LID models therefore become important factors, influencing the quality of mined text. For long tail languages, LID model performance is often degraded due to the paucity of training data in those languages. Additionally, when identifying minority languages in large web-crawled datasets, the quality of the data itself matters and investigations into the classes of *dataset noise* become of paramount importance (Caswell et al., 2020).

This paper investigates one type of noise in Cyrillic script text that can also potentially serve as a marker for low-resource orthographies: confusable characters. We consider the writing systems from the North Caucasus used to write over thirty languages from the Nakh-Daghestanian and Abkhaz-Adyge language families (Daniel and Lander, 2011; Chumakina, 2017), which we describe in Section 2. In this paper, we choose to focus on arguably the most salient common orthographic feature[1] of these writing systems — the Cyrillic let-

[1]Other letters and their confusable variants may also provide useful signals. For example, standalone *breve* (U+02D8) is found in the PanLex Swadesh list for the endangered Rutul language (Kamholz et al., 2014). Originally an element of Abkhaz orthography it is now obsolete but can be found in web crawls. We defer investigation of this and other such phenomena to future work.



*Proceedings of the Third Workshop on Computation and Written Language (CAWL 2026) @LREC 2026*, pages 50–63
12 May, 2026. 

ter *palochka* (U+04C0) and its lower-case variant.[2] This letter was introduced to the Unicode standard in 2006. Prior to that, it was often represented by the capital Latin letter *I* or the digit *1* in keyboards used to type the Cyrillic orthographies of North Caucasus languages, such as Abaza (Koshevoy et al., 2023). As we show in Section 3, several additional confusable characters are actively used as proxies for *palochka*, most likely due to the scarcity of native input methods. These and other similar confusable characters have been cited as adversely affecting the quality of LID and machine translation models for long-tail languages (Caswell et al., 2020; Bapna et al., 2022).

Our experiments in Section 4 investigate the following research questions:

*RQ1*: How sensitive are modern, wide coverage LID models to the noise introduced by confusable characters? For this purpose, we focus on the GlotLID v3 model, which is trained to identify over 2,000 language-script labels and covers 14 languages belonging to the language families of interest (Kargaran et al., 2023).[3] The answer to this question is not obvious because within this group of 14 vulnerable or endangered languages, we expect a strong *a priori* imbalance in the amount of labeled training data available to the LID model.

*RQ2*: Can confusable character normalization increase the recall for languages in question? In other words, will orthographic noise reduction lead to the discovery of new data in these languages?

*RQ3*: How sensitive is LID model precision for endangered North Caucasian languages to confusable character normalization? Can the cases when the model stops predicting one of the languages in question post-normalization be considered a reduction in the false positive rate, or an increasing of the number of false negatives?

Since there is no hand-labeled test set available for the experiments described above, we rely on the metadata of the original web documents where possible, as well as manual curation.

## 2. Background

**Cyrillic Script** From the early days in the 1960s, the computing standards for representing Cyrillic character sets have co-evolved with the Latin script standards from initial 7-bit and later 8-bit representations. This has been attributed to the similarity and relative simplicity of the two scripts as well as a considerable degree of cooperation between various international standardization bodies at the time (Ross, 1984; Hopkinson, 1984; Clews, 1988). This gradual evolution led to the first 16-bit versions of the Unicode standard being able to encode major Cyrillic writing systems that included Bulgarian, Belarusian, Macedonian, Russian, Serbian and Ukrainian (Parry, 1991; Clews, 1991; Haralambous, 2007). Beyond these and other well-resourced languages that also include the major languages of Russian Federation and Central Asia, such as Kazakh, Kyrgyz, Mongolian, Tatar and Tajik, there is a rich and linguistically well-explored orthographic landscape of many more lesser-resourced languages utilizing Cyrillic script (Musajev, 1965). Modern estimates place the number of living languages within the Russian Federation alone at somewhere between 150 (Alpatov, 2000) and 155 (Koryakov et al., 2022), the majority of which are in various states of endangerment (Kraeva and Guermanova, 2020; Gruzdeva, 2022). At the same time, the Cyrillic portion of the Unicode standard has been continuously evolving to accommodate some of these orthographies, both living and historic (Miller, 2021; Manulov, 2022; Anderson, 2023), with some writing systems still unaddressed (Roncero, 2021).

**Writing Systems of North Caucasus** We investigate the languages from Abkhaz-Adyge (Northwest Caucasian, or NWC)[4] and Nakh-Daghestanian (Northeast Caucasian, or NEC)[5] language families. These two language families encompass languages with typologically idiosyncratic features of phonology, morphology and syntax that make them an important subject of linguistic studies, with some researchers calling North Caucasus "Europe's linguistically most exotic area" (Daniel and Lander, 2011). For example, according to Chirikba (2016), the phonemic inventory of Ubykh, a recently extinct NWC language, has close to eighty consonants while only possessing three vowels, while Babaliyeva (2023) mentions 44 grammatical cases in Tabasaran, an NEC language.[6]

There are presently 37 ISO 639-3 language codes allocated to the languages from these two families, shown in Table 1. Five languages are in the NWC family and the remaining languages are in the NEC family. Of these, the NWC Ubykh and NEC Aghwan are now extinct. Despite their endangered status, there is plausibly still non-negligible amounts of text in these languages present on the

[2] The official orthography of Abkhaz is the exception in not employing *palochka* (Hewitt, 2010). However, as we show in Section 4, paragraph-level LID can assign spurious Abkhaz labels to such data.

[3] https://github.com/cisnlp/GlotLID

[4] https://glottolog.org/resource/languoid/id/abkh1242

[5] https://glottolog.org/resource/languoid/id/nakh1245

[6] This traditional view of the case system in NEC languages is contested by Comrie and Polinsky (1998).

| Code | Name | Family | Code | Name | Family |
|---|---|---|---|---|---|
| abk | Abkhaz | NWC | inh | Ingush | NEC |
| abq | Abaza | NWC | kap | Bezhta | NEC |
| ady | Adyghe | NWC | kbd | Kabardian | NWC |
| agx | Aghul | NEC | khv | Khwarshi | NEC |
| akv | Akhvakh | NEC | kjj | Khinalugh | NEC |
| ani | Andi | NEC | kpt | Karata | NEC |
| aqc | Archi | NEC | kry | Kryz | NEC |
| ava | Avar | NEC | kva | Bagvalal | NEC |
| bbl | Bats | NEC | lbe | Lak | NEC |
| bdk | Budukh | NEC | lez | Lezgian | NEC |
| bph | Botlikh | NEC | rut | Rutul | NEC |
| che | Chechen | NEC | tab | Tabasaran | NEC |
| cji | Chamalal | NEC | tin | Tindi | NEC |
| dar | Dargwa | NEC | tkr | Tsakhur | NEC |
| ddo | Tsez | NEC | uby | Ubykh | NWC |
| gdo | Godoberi | NEC | udi | Udi | NEC |
| gin | Hinuq | NEC | ugh | Kubachi | NEC |
| huz | Hunzib | NEC | xag | Aghwan | NEC |
| ... | ... | ... | xdq | Kajtak | NEC |

Table 1: The NEC and NWC languages of interest classified as either *vulnerable* (default), *definitely endangered* (blue), *severely endangered* (orange) or *extinct* (red) by UNESCO (Moseley, 2010).

Web (for example, in the form of lexicographic dictionaries), that either get filtered out by crawlers or misclassified by LID filters, due to poor support of NEC and NWC languages by current LID models.

The NEC and NWC alphabets with official status in the Russian Federation use the Cyrillic script. As we mentioned in the Introduction, the presence of the *palochka* letter is typically indicative of orthographies of these languages, where depending on the language, it modifies the preceding consonant, marking it as ejective or pharyngeal, or serves on its own as a glottal stop (Daniel and Lander, 2011). Since usage as a modifier letter is very common, the *palochka* letter is found in many digraph and trigraph letters of NEC and NWC writing systems that possess very complex consonantal inventories.[7] For example, in Kubachi orthography, the native word for the Kubachi language is written "ГIугъбуган" and the word for "hen" is "гIягIя", upper- and lower-case examples of the standalone digraph "гI" that includes *palochka* and represents the [ʔ] pharyngeal plosive.

The NEC family also includes unwritten languages. This situation is changing since new Cyrillic script-based orthographic proposals for previously unwritten languages have constantly emerged in Dagestan and beyond since the late 1990s (Ataev, 2015). Additional complexities arise when unwritten languages are spoken outside of the predominantly Cyrillic script area. For example, the NEC language Bats (closely related to Chechen and Ingush) is spoken across the border in Georgia and historically has been heavily influenced by the Georgian language of the unrelated Kartvelian family (Gippert, 2008). The proposed Bats writing systems therefore include ones based on Georgian, Cyrillic and even Latin alphabets. Similarly, tentative Latin and Cyrillic script-based orthographies were mentioned for the NEC language Kryz from a Lezgic branch spoken in Azerbaijan (Clifton et al., 2005). In addition to *digraphia*, the written material in such languages may borrow the orthography of a larger language spoken in the area, which further complicates mining for such data.

[7] The Kabardinian Cyrillic orthography includes a single tetragraph "кхъу" representing [qχʷ] sound, while the recently proposed orthography for Archi contains the tetragraph "ххьI" representing [χːˤ] (Chumakina et al., 2007).

| Name | Unicode | Visual |
|---|---|---|
| CYRILLIC CAPITAL LETTER BU I | U+0406 | І |
| CYRILLIC SMALL LETTER BU I | U+0456 | і |
| DIGIT ONE | U+0031 | 1 |
| GREEK SMALL LETTER IOTA | U+03B9 | ι |
| LATIN CAPITAL LETTER I | U+0049 | I |
| LATIN SMALL LETTER I | U+0069 | i |
| LATIN SMALL LETTER IOTA | U+0269 | ɩ |
| LATIN SMALL LETTER L | U+006C | l |

Table 2: The set of confusable character Unicode code points and corresponding visual form. We abbreviate BYELORUSSIAN-UKRAINIAN as BU to preserve space.

**Confusable Characters** Due to the visual similarity between some characters of Latin and Cyrillic scripts, these confusable characters often appear in web-crawled data. The security threats due to these characters, or *homoglyphs*, are actively studied in the cybersecurity field (Gabrilovich and Gontmakher, 2002; Holgers et al., 2006; Deng et al., 2020; Wang et al., 2024). More recently, the effects of confusable characters on the performance of large language models (LLMs) have also been investigated in the context of preventing the misuse of AI-generated text (Boucher et al., 2022; Kirchenbauer et al., 2023; Creo and Pudasaini, 2025; Cooper et al., 2025). In the next section we explore the various confusable characters used as a proxy for *palochka* in a large web-crawled corpus, and describe our initial investigation of LID for NEC and NWC languages.

## 3. Preliminary Glance at the Data

For this preliminary investigation we chose to search for the NEC and NWC languages using *palochka* confusable characters as a marker in MADLAD-400, which is a general domain multilingual dataset based on CommonCrawl that spans 419 languages (Kudugunta et al., 2023). The dataset has two partitions: a 5 trillion token `noisy` dataset, which is the dataset obtained after applying document-level LID but before any filtering, and a 3 trillion token `clean` dataset, which has a wider

variety of filters applied, including some based on the document-level LID. Since we are interested in languages mostly outside the set of MADLAD-400 LID labels, we chose to explore the `noisy` partition for our experiment. Furthermore, in our analysis in this section we follow a manual paragraph-level LID process for NEC and NWC languages using the source document metadata, online lexicographic dictionaries and similar contextual clues, where possible.

**Palochka Variants** Table 2 presents the set of eight characters that in our experiments serve as a proxy for *palochka*, lower- and upper-case variants. It consists of diverse characters with varying degrees of visual confusability with the originals. The set was assembled after inspecting various data sources. Two characters in the list (the two variants of *Latin letter I*) belong to the set of genuine homoglyphs maintained by the Unicode Consortium (2025).[8] These characters along with the *Digit 1* letter are still widely found in NEC and NWC language web pages, typically but not exclusively originating from the autonomous republics of the Russian Federation where these languages are spoken. We also included rarer characters such as Greek and Latin variants of letter *Iota* that can be found in electronic lexicographic dictionaries and Swadesh lists for a multitude of NEC and NWC languages such as PanLex (Kamholz et al., 2014).[9]

**Basic Method** We employ a simple pipeline consisting of two steps: The first step splits MADLAD-400 documents from the `noisy` set into paragraphs using newline separators. The second step selects the relevant paragraphs. We denote the set of confusable variants of *palochka* described above $Y$. In order to retrieve paragraphs that *possibly* belong to the languages of interest recorded using input methods that employ elements of $Y$, we searched for paragraphs that contain at least one token consisting entirely of Cyrillic script letters combined with the elements of $Y$. Concretely, such tokens should contain at least one sub-string $(x_1, y, x_2)$, $x_i \in X$ and $y \in Y$, where $X$ is the set of all Unicode Cyrillic *lowercase* letters.[10]

**Resulting Signal** Splitting the `noisy` MADLAD-400 documents by newline characters and filtering to those with a *palochka* confusable character yields 446,356,076 paragraphs. As we show later in this section, only a small proportion of these paragraphs belong to the NEC/NWC languages. In addition to the low-quality data, such as spam and unnatural text, the filter triggers on valid text in unrelated languages whose orthographies permit tokens that pass the *palochka* filter. In particular, the Ukrainian, Belarusian, Kazakh and Rusyn Cyrillic writing systems because two of the *palochka* variants in Table 2 (the two versions of *Ukrainian-Belarusian Letter I*) are valid characters in their orthography. As we show in Section 4, applying a paragraph-level LID filter prunes out the majority of these original paragraphs resulting in a significantly smaller set of candidate paragraphs.

After removing the paragraphs that belong to the documents with MADLAD-400 language labels corresponding to the four orthographies resulting in false positives described above, we obtained a text sample that *likely* belongs to NEC and NWC languages, but still includes some unrelated languages as well. Some examples of sample sentences in 12 languages from this set are shown in Table 3, where the snippets from the resulting paragraphs with the confusable characters highlighted in red are shown along with the manual language assignment.[11] The two examples shown in the table belonging to Turkic languages—Bashkir and Southern Altai—highlight the false positives in our set. Neither language has the letter *palochka* in its orthography and the confusable characters in these examples are likely artifacts of inadequate input methods. In the Bashkir example, in addition to the actual false positive *palochka* instance (represented as digit one), there are two further non-orthographic characters (digits two and three) that indicate problems with the document encoding. The analysis for this case is provided in Table 4.

**Document-level LID** MADLAD-400 supports five out of the 35 living NEC and NWC languages shown in Table 1: Adyghe, Avar, Chechen, Kabardian and Tabasaran. For the snippets of NEC and NWC languages in Table 3, considering the manual language assignments in the third column as the truth, we observe that paragraphs belonging to four languages from the above set match the document-level language labels in the sixth column. For paragraphs in languages outside of the supported set, the model tended to assign the document to languages in the same family. So, for instance, Lezgin paragraph gets the Tabasaran label, which is expected as both languages reside in the Lezgic branch of NEC

[8] https://www.unicode.org/Public/security/latest/confusables.txt

[9] https://db.panlex.org/

[10] We ignore uppercase letters for left and right context in this paper. Our filter misses all the uppercase words (such as titles) from NEC and NWC text but at the same time has better precision because *palochka* variants in lowercase Cyrillic context are more likely to occur for languages of interest in general multilingual corpus.

[11] We located the source web documents and identified the languages from the context. For example, the Rutul sentence comes from bilingual Rutul-Russian newspaper https://rutnov.ru/.

| Sample Sentence | Char(s) | Manual Language Assignment | | | Document |
|---|---|---|---|---|---|
| | | Language | Code | Family | Language |
| УадыргIвана амартанква гьажьыта расацIла бацала йыршшуа йалагатI | U+0049 | Abaza | abq | NWC | Avar |
| Къэбар гухэкIыбэ къэзыхьырэ уз | U+0049 | Adyghe | ady | NWC | Adyghe |
| ХIаттарис хIуьлмат дагулурай хье арайиъ. | U+0049 | Aghul | agx | NEC | Russian |
| Бу уредуни уренерге болушкан улус коркуш коп оны чотозо, чазын 1етпес. Ончогорго1аан быйан. | U+0031 | Southern Altai | alt | Turkic | Southern Altai |
| Салатавиялъул диалекталъулъ туркази лъарагI рагIаби | U+0049 | Avar | ava | NEC | Avar |
| УЗытыу тыу1ан телдә алып барылған мәктәптәр2ең 2-се класы өсөн баш3орт (дәүләт) теленән эш программаһы. | U+0031 | Bashkir | bak | Turkic | Bashkir |
| Iин чуьрчу боданехь таьIна сайн декхнаш а гуш! | U+0049 | Chechen | che | NEC | Chechen |
| Ва пагьмучерси дарган, мухбир хIед аррукири? | U+0049 | Dargwa | dar | NEC | Avar |
| Берд Юрт яха оагIув укхазахьа хьожаяьй | U+0049 | Ingush | inh | NEC | Chechen |
| ЩIэныгъэлIхэр игъащIэ лъандэрэ топсэлъыхь цIыхумрэ ар къэзыухъуреихь дунеймрэ я псэкIэ | U+0049 | Kabardian | kbd | NWC | Kabardian |
| Гьинтнил кIиришиву махъ лирчIсса, кIинтнил замгьарду байбишин бувасса ссутнил гьантри хас бувара зула ... | U+0049 | Lak | lbe | NEC | Avar |
| Нугъатрин, рахунрин кьет1енвилериз килигна лезги ч1ал пуд наречиедиз пай жезва | U+0031 | Lezgin | lez | NEC | Tabasaran |
| Гьабише Аллагьахьде дилег гьаъара, джва1ршиклаа гигубыр гьаъ, хьур. | U+0031 | Rutul | rut | NEC | Avar |
| ХанаIгьмад чоджий, ЦIаIхни хивын джегьилер. | U+0049 | Tsakhur | tkr | NEC | Avar |

Table 3: Sample of MADLAD-400 Cyrillic sentences in different orthographies with highlighted *palochka* confusables. The twelve NEC and NWC manually identified languages shown here are classified as either definitely endangered or vulnerable. The MADLAD-400 document language labels are shown in sixth column.

| Original text | Corrected text |
|---|---|
| БАШҒОРТОСТАН РЕСПУБЛИКА*ЫНЫ4М(!АРИФ МИНИСТРЛЫ!Ы | БАШҒОРТОСТАН РЕСПУБЛИКАҺЫНЫҢ МӘҒАРИФ МИНИСТРЛЫҒЫ |
| МР Й(РМ(К(Й РАЙОНЫНЫ4 М(!АРИФ ИДАРАЛЫ!Ы | МР ЙӘРМӘКӘЙ РАЙОНЫНЫҢ МӘҒАРИФ ИДАРАЛЫҒЫ |
| ТАРҒКА2Ы АУЫЛЫ УРТА БЕЛЕМ БИРЕҮ М(КТ(БЕ | ТАРҒКАҙЫ АУЫЛЫ УРТА БЕЛЕМ БИРЕҮ МӘКТӘБЕ |
| Ф(ННИ- ТИКШЕРЕНЕҮЭШЕ | ФӘННИ- ТИКШЕРЕНЕҮЭШЕ |
| Тарка2ы хал3ы телм9ренд9 фразеологизмдар2ы 3улланыу. | Таркаҙы халҡы телмәрендә фразеологизмдарҙы ҡулланыу. |
| Номинация “Баш3орт теле 89м 929би9те” | Номинация “Башҡорт теле һәм әҙәбиәте” |

Table 4: Snippet of badly encoded Bashkir text from an essay in Microsoft Word*. The corrupted glyphs are highlighted in red on the left-hand side of the table, the hand-corrected versions are highlighted in blue on the right-hand side. *Source: https://nsportal.ru/ap/library/drugoe/2019/01/06/ispolzovanie-frazeologizmov-v-rechi-zhiteley-s-tarkazy.

family, while Ingush gets labeled as Chechen. The behavior of the model, however, is more nuanced because the assignment to the closest language is not always followed. For example, Lak and Dargwa paragraphs get labeled as Avar – all NEC languages, but not closely related. Another example is the paragraph in Tsakhur, which is also labeled as Avar even though one might expect Tabasaran, since it is closer phylogenetically. In a more problematic scenario, some sentences get labels from different language families, e.g., the snippet in Abaza, an NWC language, is labeled as Avar from the NEC family.

Identifying the out-of-domain languages and estimating the amount of available language data is difficult because in the absence of a reliable paragraph-level LID model with enough coverage, the manual process described above is not scalable. In the next section, we investigate improvements to our pipeline involving the incorporation of a paragraph-level LID filter with wider coverage of NEC and NWC languages, and study the behavior of such a paragraph-level model in the presence of *palochka* variants.

## 4. Experiments and Discussion

The following analysis examines the interplay between the *palochka* variant characters and an LID model. Unlike the prior mostly manual inspection procedure to check for the presence of useful NEC and NWC language signals described in Section 3, our method here uses a wide coverage document- and paragraph-level LID model and examines a significantly bigger corpus.

**The Dataset and LID Model** For the main body of experiments we use DCAD-2000 (Shen et al., 2025) — a recently introduced large-scale multilingual corpus which is significantly larger than MADLAD-400, contains more recent crawls from CommonCrawl and includes the other multilingual sources such as FineWeb-2 (Penedo et al., 2025) and MaLA (Lin et al., 2024). Crucially, the dataset covers 2,282 document-level language-script LID labels produced by GlotLID (Kargaran et al., 2023),

which is a LID model using fastText architecture (Joulin et al., 2016). The subset of NEC and NWC languages supported by GlotLID consists of 14 Cyrillic writing systems. There are four NWC languages — Abaza (abq), Abkhaz (abk), Adyghe (ady), Kabardian (kbd) — and ten NEC languages — Aghul (agx), Avar (ava), Bezhta (kap), Chechen (che), Dargwa (dar), Ingush (inh), Lak (lbe), Lezgin (lez), Tabasaran (tab) and Tsakhur (tkr). This LID model has the widest NEC and NWC language coverage known to us, and thus we choose it as a default model in our experiments. We use the default GlotLID model, and set $k = 5$ for top-$k$ label extraction with a minimum hypothesis probability of $0.01$ in the model decoding hyperparameters.

Similar to experiments in Section 3 where the noisy MADLAD-400 subset was used, below we operate on the unfiltered union of keep and remove DCAD-2000 subsets. We hypothesize that the set of documents discarded by DCAD-2000 filters contains valuable data from NEC and NWC languages. Some of these documents may have been filtered by mistake, as the presence of confusable letters may have caused the removal filters to trigger, but also possibly due to poor LID.

**Gauging the Amount of Useful Signal** Applying the *palochka* confusables filter from Section 3 to DCAD-2000 data yields 762,639,585 paragraphs of text in various writing systems that contain at least one Cyrillic token with *palochka* variants, which is nearly double the number of paragraphs extracted using the same technique from the MADLAD-400 corpus in Section 3. This set includes 1,574,518 paragraphs where the majority script of that paragraph's constituent tokens is not Cyrillic. We filter out these paragraphs, which include text in 156 unique scripts (including the ISO 15924 script code Zzzz for *uncoded script*). We further filter the resulting set to exclude 15,129,549 short snippets consisting of less than three tokens, since such short sequences adversely affect LID.

According to paragraph-level GlotLID predictions 98.7% of the remaining 745,935,518 paragraphs belong with high confidence to Cyrillic writing systems where the *palochka* filter triggers on valid orthographic tokens. These writing systems are the relatively high-resource Belarusian, Kazakh, Ukrainian and (low-resource) Rusyn orthographies mentioned in Section 3. After excluding these paragraphs, the remaining 9,433,217 paragraphs contain some proportion of paragraphs in NEC and NWC languages that include tokens with *palochka* letter variants. In addition, this set also includes paragraphs in unrelated Cyrillic writing systems that include false positives due to typos, bad data encoding, inadequate input methods or malicious content as demonstrated by the Bashkir example in Table 4. Finally, this also includes some instances of Belarusian, Kazakh, Ukrainian and Rusyn that have been mis-classified as other languages.

The paragraph-level GlotLID predictions for NEC and NWC languages include 1,628,526 paragraphs in 14 languages as shown in the first and second columns of Table 5. Among the remaining 7,799,786 paragraphs in unrelated languages the top five high-frequency entries in the language list are Russian (rus), "Undetermined" (und), Old Russian (orv), Komi-Zyrian (kpv) and Bulgarian (bul) Cyrillic orthographies, where "Undetermined" refers to cases when GlotLID failed to assign a label. The *palochka* letter is not part of the last three writing systems in the list and does not feature in the orthography of modern Russian, but may be present in Russian paragraphs that feature NEC or NWC language code-switching. Additional signal may potentially be found among paragraphs with "Undetermined" label.

**Filter Precision and Recall** The application of LID after the *palochka* variant filter serves to separate the true positives (1,628,526 paragraphs) from the false positives in unrelated languages (7,799,786 paragraphs), yielding the *palochka* variant filter precision of 17.3%. Computing the corresponding filter recall requires a different pipeline, where LID is applied before the variant filter. This recall-specific pipeline yields a comparable number of true positives and significantly higher number of false negatives resulting in a low recall value of 7.4%.[12] The detailed analysis of the recall pipeline is provided in Appendix A.

**Normalization of *Palochka* Variants** The normalization pipeline is a straightforward extension of the *palochka* variant filters from Section 3: the input text is split into paragraphs, followed by the search for *palochka* variants in Cyrillic tokens. The additional step required for normalization involves the context-dependent rewrite rule that maps these *palochka* variants $y$ from the set of variants $Y$ (from Table 2) surrounded by lowercase Cyrillic letters to some canonical representation $\hat{y}$, which we set to *Cyrillic Small Letter Palochka* (U+04CF).

The differences in paragraph-level best hypothesis GlotLID performance before and after normalization of *palochka* letter variants are shown in Ta-

[12]The high number of false negatives in this scenario is not surprising. If the LID prediction is indeed correct it is not necessarily the case that the particular NEC/NWC language paragraph in question must necessarily use letter *palochka* or its variants. For example, as we mentioned earlier, there are not *supposed* to be instances of *palochka* in Abkhaz orthography, yet we observe a small number of such instances.

| Lang. | Paragraphs | | Change (%) | Wins | Losses | Identity | Same Group | Lengths | |
|---|---|---|---|---|---|---|---|---|---|
| | Before | After | | | | | | $\mu$ | $\sigma^2$ |
| abk | 6008 | 6340 | 5.237 | 1118 | 930 | 4210 | 5078 | 4.25 | 50.71 |
| abq | 14 326 | 11 676 | −18.498 | 11 | 854 | 11 646 | 13 472 | 28.2 | 741.25 |
| ady | 136 143 | 136 359 | 0.158 | 148 | 112 | 131 119 | 136 031 | 33.09 | 28 222.37 |
| agx | 8008 | 7462 | −6.818 | 39 | 112 | 7258 | 7896 | 33.79 | 1473.21 |
| ava | 248 123 | 257 840 | 3.769 | 4261 | 111 | 247 990 | 248 012 | 33.92 | 1425.24 |
| che | 567 820 | 564 611 | −0.565 | 539 | 2497 | 563 142 | 565 323 | 36.72 | 2330.68 |
| dar | 21 023 | 20 822 | −0.956 | 135 | 153 | 20 551 | 20 870 | 36.2 | 989.31 |
| inh | 76 985 | 75 579 | −1.826 | 883 | 2054 | 73 856 | 74 931 | 108.33 | 164 566.24 |
| kap | 534 | 167 | −68.727 | 15 | 21 | 137 | 513 | 6.87 | 247.65 |
| kbd | 396 096 | 397 616 | 0.382 | 2407 | 736 | 390 027 | 395 360 | 27.51 | 1043.34 |
| lbe | 46 895 | 47 248 | 0.747 | 537 | 253 | 46 065 | 46 642 | 35.99 | 1629.85 |
| lez | 72 183 | 70 204 | −2.742 | 304 | 627 | 69 415 | 71 556 | 41.71 | 2495.88 |
| tab | 33 448 | 33 754 | 0.907 | 255 | 151 | 32 722 | 33 297 | 40.94 | 1171.15 |
| tkr | 934 | 884 | −5.353 | 16 | 21 | 807 | 913 | 26.98 | 729.09 |
| All | 1 628 526 | 1 630 562 | 0.125 | 10 668 | 8632 | 1 598 945 | 1 619 894 | 37.18 | 11 927.56 |

Table 5: Summary of top LID prediction differences for NEC and NWC languages in DCAD-2000 before and after the normalization of *palochka* variants.

| Lang. | Size | Before | | | | After | | | | $t$-test | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | $N$ | $\mu$ | $\sigma^2$ | CI$^-$ | CI$^+$ | $\mu$ | $\sigma^2$ | CI$^-$ | CI$^+$ | $t$-stat | $p$-val |
| abk | 4210 | 0.487 | 0.093 | 0.478 | 0.496 | 0.521 | 0.093 | 0.512 | 0.530 | −5.133 | 2.908e−7 |
| abq | 11 646 | 0.964 | 0.010 | 0.962 | 0.966 | 0.798 | 0.061 | 0.793 | 0.802 | 67.500 | 0.0 |
| ady | 131 119 | 0.759 | 0.066 | 0.758 | 0.761 | 0.759 | 0.065 | 0.757 | 0.760 | 0.641 | 0.521 |
| agx | 7258 | 0.878 | 0.056 | 0.872 | 0.883 | 0.863 | 0.064 | 0.857 | 0.869 | 3.571 | 0.0 |
| ava | 247 990 | 0.961 | 0.020 | 0.960 | 0.961 | 0.990 | 0.004 | 0.989 | 0.990 | −93.649 | 0.0 |
| che | 563 142 | 0.984 | 0.008 | 0.984 | 0.984 | 0.981 | 0.010 | 0.981 | 0.981 | 18.741 | 2.338e−78 |
| dar | 20 551 | 0.857 | 0.045 | 0.854 | 0.859 | 0.851 | 0.046 | 0.848 | 0.854 | 2.491 | 0.013 |
| inh | 73 856 | 0.869 | 0.064 | 0.867 | 0.871 | 0.866 | 0.066 | 0.864 | 0.868 | 2.273 | 0.023 |
| kap | 137 | 0.544 | 0.095 | 0.492 | 0.596 | 0.489 | 0.080 | 0.441 | 0.537 | 1.542 | 0.124 |
| kbd | 390 027 | 0.935 | 0.032 | 0.934 | 0.935 | 0.940 | 0.031 | 0.940 | 0.941 | −13.123 | 2.469e−39 |
| lbe | 46 065 | 0.851 | 0.047 | 0.849 | 0.853 | 0.857 | 0.044 | 0.856 | 0.859 | −4.565 | 4.999e−6 |
| lez | 69 415 | 0.889 | 0.036 | 0.887 | 0.890 | 0.875 | 0.042 | 0.874 | 0.877 | 12.792 | 1.899e−37 |
| tab | 32 722 | 0.929 | 0.039 | 0.927 | 0.932 | 0.942 | 0.032 | 0.940 | 0.943 | −8.313 | 9.522e−17 |
| tkr | 807 | 0.629 | 0.101 | 0.607 | 0.651 | 0.636 | 0.098 | 0.614 | 0.657 | −0.419 | 0.675 |

Table 6: Differences in top LID prediction confidences before and after normalization for the cases when the same language is predicted. The second column indicates the size of the population ($N$) that corresponds to the seventh column named “Identity” in Table 5. The cases when normalization results in increased prediction confidences are marked in blue, the confidence degradations are marked in red.

ble 5. For each language the number of paragraphs is shown before and after applying the normalization. The percentage increase or decrease in the paragraph count is indicated in the “Change” column. The counts for the cases when the top post-normalization prediction changes from the unrelated language to the language in NEC or NWC families (e.g., from Russian to Chechen) are provided in the “Wins” column. Conversely, the counts of cases when the top pre-normalization prediction switches from an NEC or NWC language to a language from an unrelated family (e.g., from Kabardian to Karachay–Balkar) are indicated in “Losses” column. The column “Identity” provides the counts for the cases when top language prediction is not affected by normalization. The “Same Group” column provides accumulated counts for the cases when both the pre-normalization and post-normalization top hypothesis belongs to either NEC or NWC language family (e.g., Avar changes to Ingush). Finally, the “Lengths” column provides statistics on paragraph lengths.

As can be seen from Table 5, overall the effects of normalization are positive but minor resulting in just 2,036 new sentences for the 14 languages in question. These effects vary by language. According to GlotLID more paragraphs being discovered after the normalization in six languages (Abkhaz, Adyghe, Avar, Kabardian, Lak and Tabasaran), while some paragraphs are lost for another eight (Abaza, Aghul, Bezhta, Chechen, Dargwa, Ingush, Lezgin and Tsakhur). We note that very little data in Bezhta and Tsakhur make it through our filters, which reflects their limited coverage in DCAD-2000.

We next investigate the effect of normalization on prediction confidence for the languages of interest when the best prediction is not changed by normalization for which the relevant counts are shown in column "Identity" of Table 5. We compute the best prediction confidence as a difference between the best and second-best hypothesis. If normalization has positive effect on the overall LID performance one would expect an increase in confidence; conversely, a decrease in confidence points at model confusion. The differences in top prediction confidences and the relevant statistics are shown in Table 6. For each language, for a sample size of $N$ paragraphs we compute the population mean ($\mu$), variance ($\sigma^2$) and 95% confidence intervals ($[\mathrm{CI}^-, \mathrm{CI}^+]$ for the sets of prediction confidences before and after the normalization. To validate whether the per-language prediction confidence differences are statistically significant we perform a two-sample $t$-test on the two populations and provide the relevant $t$-stat score and $p$-value. The null hypothesis rejection threshold is set to $0.05$. The $t$-test results indicating an overall improvement in confidence after the normalization are marked in blue, the cases when normalization harms the confidence are marked in red, while the rest of the cases indicate that normalization has no effect.

According to Table 6, the set of languages for which prediction confidence improvements are observed coincides with the set of languages in Table 5 for which more data is discovered post-normalization with the exception of Adyghe where normalization has no effect on same-language prediction confidence. Similarly, the set of languages where the post-normalization confidence degrades correlates with the languages with negative post-normalization effects in Table 5 with the exception of Bezhta and Tsakhur where no significant changes in prediction confidence are observed likely due to small population size, and Dargwa and Ingush, where the drop in confidence is significant but relatively small.

**Inspection of the Differences** We perform a manual inspection of randomly sampled 140 paragraphs (10 for each language) gained or lost during the normalization. The results are shown in Table 7, where for each "win" or "loss" category we maintain "exact" ($N_e$) and "lenient" ($N_l$) counters. The "exact" counter represents the exact matches between manually established label and the GlotLID label. Some Tsakhur paragraphs, for example, are actually in Rutul according to our inspection. In a more permissive "lenient" mode, we still count this mismatch as a match because Rutul is not supported by GlotLID but belongs to languages of interest (see Table 1). In the exact mode, we treat this mismatch as error. As

| Lang. | Wins | | Losses | | Lang. | Wins | | Losses | |
|---|---|---|---|---|---|---|---|---|---|
| | $N_e$ | $N_l$ | $N_e$ | $N_l$ | | $N_e$ | $N_l$ | $N_e$ | $N_l$ |
| abk | 0 | 0 | 0 | 0 | inh | 0 | 0 | 2 | 2 |
| abq | 3 | 4 | 8 | 8 | kap | 2 | 2 | 0 | 0 |
| ady | 8 | 8 | 9 | 9 | kbd | 6 | 6 | 7 | 7 |
| agx | 2 | 3 | 0 | 1 | lbe | 4 | 4 | 7 | 7 |
| ava | 9 | 9 | 9 | 9 | lez | 7 | 7 | 7 | 7 |
| che | 0 | 0 | 3 | 4 | tab | 3 | 4 | 1 | 1 |
| dar | 10 | 10 | 10 | 10 | tkr | 4 | 8 | 1 | 3 |
| ... | ... | ... | ... | ... | All | 58 | 65 | 64 | 68 |

Table 7: Manual validation of GlotLID language assignments on a random 10-paragraph sample in "wins" and "losses" categories for each language. Two modes of counting are employed: exact ($N_e$) and lenient ($N_l$).

can be seen from the table, for "wins" category if counted exactly only 58 (or 41% of 140 new paragraphs) can be counted as real improvements in recall after normalization. The rest of the paragraphs are false positives due to GlotLID misclassifications. For example, we found no Abkhaz, Chechen or Ingush data in the sampled "win" category because all the "Abkhaz" paragraphs are in Ukrainian, "Chechen" paragraphs are random itemized lists of place names and Ingush data, whereas the "Ingush" label gets assigned to mostly Chechen data.

When evaluating the losses we check that pre-normalization language assigned by GlotLID is correct. If it is not, then the paragraph signal was wrong and there is nothing to declare as a loss. For example, for a sample in Table 7 normalization harms a perfectly valid Abaza sentence "Ужвы датша йыгIгIвуаш ажвакI хIвастI" which gets a Belarusian GlotLID label after normalization. For Bezhta, on the other hand, none of the source sentences are in Bezhta (most of them are in a higher-resource neighboring Avar) hence none of the data is truly lost. The inspection of the losses also reveals signals from NEC languages unsupported by GlotLID. For example, the sentences "ОъшIаъ доьвен са нишанне тастIа гъара ехне" and "Ед' ши экIра, экIра инджимишь гьеъа" in severely endangered Udi and definitely endangered Botlikh (Table 1) are labeled by GlotLID as Aghul and Tsakhur, respectively. We also note relatively high overall agreement between manual and GlotLID labels for Adyghe, Avar and Dargwa. According to Table 7 only 46% of "lost" paragraphs can be considered as losses if counted "exactly".

The only language with 100% disagreement between the automatic GlotLID and manually assigned labels is Abkhaz. Further inspection reveals that none of the paragraphs labeled as Abkhaz are in that language. Since these paragraphs include characters from *palochka* confusable set one would expect a label from another member

of NWC family where these characters are legal. However most of these paragraphs are in languages outside the NWC or NEC families, which is likely due to the abnormal (relative to other languages) distribution of paragraph lengths in "Abkhaz" shown in Table 5, which indicates very short paragraphs ($4.25$ tokens on average) with high consistency (low variance). These paragraphs may be too short for GlotLID to identify correctly.[13]

## 5. Conclusions

We have investigated the utility of string tokens with *palochka* confusable letter variants for mining multiple writing systems that share that letter. Both the manual inspection and filtering by paragraph-level LID revealed useful NEC and NWC language signals, in particular in languages currently unsupported by the off-the-shelf large coverage GlotLID model. We also explored the effects of confusable character normalization on total recall (*RQ2*) and precision (*RQ3*), which were found to be highly language-specific. We manually evaluated the differences in GlotLID assignments on a small sample of paragraphs and discovered that predictions are mostly unreliable for nine languages out of 14. The error analysis includes data quality issues and language confusions due to phylogenetic similarities leading us to conclude that GlotLID model is sensitive to confusable character noise for this set of languages (*RQ1*). This evaluation revealed further useful signals from out-of-domain languages, such as Udi and Botlikh, in the resulting data.

## 6. Limitations and Future Work

This exploratory work has several important limitations. Because none of the authors are native speakers of North Caucasian languages, the manual validation of candidate paragraphs and their LID assignments was both time-consuming and brittle. This step relied heavily on existing document metadata and online lexical resources, which are occasionally encoded in legacy orthographies. To improve upon these experiments and facilitate future research, a critical next step is the creation of a public-domain human-annotated, paragraph-level evaluation dataset. This dataset should provide broad coverage of NWC and NEC languages, capturing the orthographic variation of the kind observed in this paper. Crucially, such annotations must account for code-switching, as texts from this region naturally exhibit frequent mixing between languages. Beyond evaluation, this resource would be highly valuable for LLM-driven development, providing necessary data for few-shot prompting and model fine-tuning.

This work relied on two LID sources—document-level labels in MADLAD-400 from a proprietary model, and predictions from the wider-coverage GlotLID—to investigate whether LID can aid the discovery of useful language data that might otherwise be lost without orthographic normalization. While other state-of-the-art LID models like OpenLID (Burchell et al., 2023) exist, they currently lack coverage for NEC and NWC languages. Our reliance on these off-the-shelf models may partially explain why our observed improvements are marginal, as aptly observed by reviewers, and highly language-specific; the long-tail, low-resource languages in our study are simply not prioritized by current state-of-the-art LID systems. Consequently, improving the LID models themselves to better serve data discovery remains an important direction for future research. More specifically, a crucial direction for future work is determining whether to normalize the LID training data, retain *palochka* orthographic variants unaltered, or intentionally inject additional orthographic variance in to improve the robustness of the resulting LID models. Careful investigation of the data that goes into training of LID models and data selection strategies resulting in better model features, especially relevant to NEC and NWC languages, are also crucial. One such model-centric program is outlined in Appendix B.

Another promising research direction is investigating LID strategies tailored specifically to highly data-scarce languages. These approaches could leverage native lexicons (Selamat and Akosu, 2016; Duvenhage, 2019) or Swadesh lists, potentially synthesizing additional training data via unsupervised morphological analyzers (Smit et al., 2014; Grönroos et al., 2020).[14]

Finally, our investigation focused exclusively on letter *palochka* and its variants, which is the iconic visual hallmark of NEC and NWC writing systems marking ejectives and glottal stops. The filters described in this study can be extended in the future to include other important orthographic features specific to these writing systems that distinguish them from the rest of Cyrillic orthographies. One example of such features are the rich consonantal clusters mentioned in Section 2: due to poor coverage of base Cyrillic letters, these languages rely heavily on digraphs, trigraphs, and even tetragraphs to represent single phonemes. This has a stacking effect—a base Cyrillic letter is combined with "modifiers" (the *palochka*, the hard sign, the soft sign, or labialization markers) to represent a specific phoneme.

[13] We observe similar "Abkhaz" paragraph length distribution in data mined from MADLAD-400.

[14] It should be noted however that even unsupervised morphological analyzers require a non-trivial amount of unannotated raw text to train reliably.

## 7. Ethics Statement

All data and models we use are publicly available. The source web-crawled datasets that we inspect (MADLAD-400 and DCAD-2000) may include the crawled sources of unknown copyright provenance, hence we are not releasing any resulting derived data. We are committed to more accessible and inclusive NLP, and hope that this exploratory work contributes to extending computational approaches to these two families of vulnerable and endangered languages hitherto largely unexplored by the NLP community.

## 8. Acknowledgements

The authors thank Lawrence Wolf-Sonkin, Vitaly Nikolaev and anonymous reviewers for many useful suggestions for improving this paper.

## 9. Bibliographical References

Vladimir M. Alpatov. 2000. *150 yazykov i politika, 1917–2000. Sotsiolingvisticheskie problemy SSSR i postsovetskogo prostranstva (*150 languages and politics, 1970–2000. Sociolinguistic problems in the USSR and post-Soviet countries*)*, 2nd edition. Kraft+, Moscow. Institute of Oriental Studies, Russian Academy of Sciences (RAN). In Russian.

Deborah Anderson. 2023. RE: Comments on CYRILLIC CHE WITH HOOK's use in Khanty and Tofa (Tofalar). ISO/IEC 10646 JTC1/SC2/WG2 L2/23-015, Unicode Consortium.

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.

Boris M. Ataev. 2015. The role of Bible translation in preserving the languages of Dagestan. In Marianne Beerle-Moor and Vitaly Voinov, editors, *Language vitality through Bible translation*, volume 95 of *Berkeley Insights in Linguistics and Semiotics*, chapter 12, pages 207–216. Peter Lang Academic Publishers.

Ayten Babaliyeva. 2023. Standard Tabasaran: short grammar sketch. *Languages of the Caucasus*, 6.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Nicholas Boucher, Ilia Shumailov, Ross Anderson, and Nicolas Papernot. 2022. Bad characters: Imperceptible NLP attacks. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1987–2004, San Francisco, CA. IEEE.

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.

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.

Vyacheslav Chirikba. 2016. From North to North West: How North-West Caucasian evolved from North Caucasian. *Mother Tongue: Journal of the Association for the Study of Language in Prehistory*, 21:1–28.

Marina Chumakina. 2017. Caucasian languages. *Oxford Research Encyclopedia of Linguistics*.

Marina Chumakina, Dunstan Brown, Greville G. Corbett, and Harely Quilliam. 2007. A dictionary of the languages of the Archi villages, south Daghestan (Archi-Russian-English). (Online edition). University of Surray, UK.

John Clews. 1988. *Language Automation Worldwide: The Development of Character Set Standards*. British Library Research & Development Reports. SESAME Computer Projects, Harrogate, UK.

John Clews. 1991. Dealing with multiple languages in the computer industry computer character sets: their evolution and impact. In *Proceedings of Translating and the Computer 13: The theory and practice of machine translation – a marriage of convenience?*, London, UK. Aslib.

John M. Clifton, Janfer Mak, Gabriela Deckinga, Laura Lucht, and Calvin Tiessen. 2005. The sociolinguistic situation of the Kryz in Azerbaijan. Survey Report SIL Electronic Survey Reports 2005–006, SIL International, Journal of Language Survey Reports.

Bernard Comrie and Maria Polinsky. 1998. The great Daghestan case coax. In Anna Siewierska and Jae Jung Song, editors, *Case, Typology and Grammar*, volume 38 of *Typological Studies in Language*, pages 95–114. John Benjamins Publishing Company.

Portia Cooper, Eduardo Blanco, and Mihai Surdeanu. 2025. The lies characters tell: Utilizing large language models to normalize adversarial Unicode perturbations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18932–18944, Vienna, Austria. Association for Computational Linguistics.

Aldan Creo and Shushanta Pudasaini. 2025. SilverSpeak: Evading AI-generated text detectors using homoglyphs. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 1–46, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Michael Daniel and Yury Lander. 2011. The Caucasian languages. In Bernd Kortmann and Johan van der Auwera, editors, *The Languages and Linguistics of Europe: A comprehensive guide*, volume 1 of *The World of Linguistics (WOL)*, pages 125–158. De Gruyter Mouton, Berlin, Germany.

Perry Deng, Cooper Linsky, and Matthew Wright. 2020. Weaponizing Unicodes with deep learning-identifying homoglyphs with weakly labeled data. In *Proceedings of 2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6, Arlington, VA. IEEE.

Bernardt Duvenhage. 2019. Short text language identification for under resourced languages. In *Proceedings of 33rd Conference on Neural Information Processing Systems (NeurIPS 2019) Workshop on Machine Learning for the Developing World: Challenges and Risks of ML4D*, pages 1–6, Vancouver, Canada.

Evgeniy Gabrilovich and Alex Gontmakher. 2002. The homograph attack. *Communications of the ACM*, 45(2):128–129.

Jost Gippert. 2008. Endangered Caucasian languages in Georgia: Linguistic parameters of language endangerment. In K. David Harrison, David S. Rood, and Arienne Dwyer, editors, *Lessons from Documented Endangered Languages*, volume 78 of *Typological Studies in Language*, pages 159–194. De Gruyter Brill.

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.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2020. Morfessor EM+Prune: Improved subword segmentation with expectation maximization and pruning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3944–3953, Marseille, France. European Language Resources Association.

Ekaterina Gruzdeva. 2022. Preserving the languages of Russia: Work in progress. *International Journal of Eurasian Linguistics*, 4(1):65–74.

Yannis Haralambous. 2007. *Fonts & Encodings*. O'Reilly Media, Sebastopol, CA.

George Hewitt. 2010. *Abkhaz: A Comprehensive Self-Tutor*. LINCOM Student Grammars. LINCOM Europa, Munich, Germany.

Tobias Holgers, David E. Watson, and Steven D. Gribble. 2006. Cutting through the confusion: a measurement study of homograph attacks. In *Proceedings of the Annual Conference on USENIX '06 Annual Technical Conference*, ATEC '06, pages 261–266, USA. USENIX Association.

Alan Hopkinson. 1984. International access to bibliographic data: MARC and MARC-related activities. *Journal of Documentation*, 40(1):13–24.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, Saule Belginova, Ariana Kenbayeva, Amina Alisheva, Abdullatif Köksal, Samir Rustamov, and Duygu Ataman. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838, Vienna, Austria. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

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.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. GlotCC: An open broad-coverage CommonCrawl corpus and pipeline for minority languages. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, Vancouver, Canada.

Erzhen V. Khilkhanova. 2019. The Internet and minority languages of Russia: Symbolic presence or a revitalization tool? (a case study of the Buryat language). *Mongolian Studies*, 11(4):967–988.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Y. B. Koryakov, T. I. Davidyuk, V. S. Kharitonov, A. P. Yevstigneyeva, and A. A. Syuryun. 2022. Spisok yazykov Rossii i statusy ikh vital'nosti (*List of languages of Russia and their vitality status*). Technical report, Institute of Linguistics, Russian Academy of Sciences (RAN), Moscow. Monograph preprint (in Russian).

Alexey Koshevoy, Anastasia Panova, and Ilya Makarchuk. 2023. Building a Universal Dependencies treebank for a polysynthetic language: the case of Abaza. In *Proceedings of the Sixth Workshop on Universal Dependencies (UDW, GURT/SyntaxFest 2023)*, pages 1–6, Washington, D.C. Association for Computational Linguistics.

Irina Kraeva and Natalia Guermanova. 2020. Language policy of the Russian Federation: Searching for balance among 150 languages. *European Journal of Language Policy*, 12(2):135–162.

Stefan Krsteski, Borjan Sazdov, Matea Tashkovska, Branislav Gerazov, and Hristijan Gjoreski. 2025. Towards open foundation language model and corpus for Macedonian: A low-resource language. In *Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025)*, pages 44–57, Vienna, Austria. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems (NeurIPS)*, 36:67284–67296.

Ali Kuzhuget, Airana Mongush, and Nachyn-Enkhedorzhu Oorzhak. 2024. Enhancing tuvan language resources through the FLORES dataset. In *Proceedings of the Ninth Conference on Machine Translation*, pages 593–599, Miami, Florida, USA. Association for Computational Linguistics.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. MaLA-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Nikita Manulov. 2022. Proposal to encode 23 Cyrillic characters for old Uslar's Caucasian Alphabets. ISO/IEC 10646 JTC1/SC2/WG2 L2/22-262, Unicode Consortium.

Kirk Miller. 2021. Unicode request for Cyrillic modifier letters. ISO/IEC 10646 JTC1/SC2/WG2 L2/21-107, Unicode Consortium.

Christopher and Moseley. 2010. *Atlas of the World's Languages in Danger*. Memory of peoples. UNESCO.

Kenesbaj Musaevič Musajev. 1965. *Alfavity jazykov narodov SSSR (*Alphabets of the languages of the peoples of the USSR*)*. USSR Academy of Sciences, Nauka, Moscow. In Russian.

E. V. Nikitina, T. N. Evgrafova, and E. I. Antonova. 2019. Russian minority languages representation on the Internet as their social status reflection. In *Proceedings of the 11th International Conference on Communicative Strategies*

*of Information Society (CSIS'19*, pages 339–348, Saint-Petersburg, Russia.

John Parry. 1991. Computer character sets: their evolution and impact. In *Proceedings of Translating and the Computer 13: The theory and practice of machine translation – a marriage of convenience?*, London, UK. Aslib.

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. *arXiv preprint arXiv:2506.20920*.

Kristian Roncero. 2021. Gakvarian Chamalal (Dagestan and Chechnya) — Language snapshot. *Language Documentation and Description*, 20:64–74.

Hugh McGregor Ross. 1984. Handling non-Roman character sets with computers. In *Proceedings of Translating and the Computer 6: Translation and communication*, London, UK. Aslib.

Jack Rueter, Olga Erina, and Nadezhda Kabaeva. 2024. On Erzya and Moksha corpora and analyzer development, ERME-PSLA 1950s. In *Proceedings of the 9th International Workshop on Computational Linguistics for Uralic Languages*, pages 67–75, Helsinki, Finland. Association for Computational Linguistics.

Ali Selamat and Nicholas Akosu. 2016. Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University - Computer and Information Sciences*, 28(4):457–469.

Yingli Shen, Wen Lai, Shuo Wang, Xueren Zhang, Kangyang Luo, Alexander Fraser, and Maosong Sun. 2025. DCAD-2000: A multilingual dataset across 2000+ languages with data cleaning as anomaly detection. *arXiv preprint arXiv:2502.11546*.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Mukhammed Togmanov, Nurdaulet Mukhituly, Diana Turmakhan, Jonibek Mansurov, Maiya Goloburda, Akhmed Sakip, Zhuohan Xie, Yuxia Wang, Bekassyl Syzdykov, Nurkhan Laiyk, Alham Fikri Aji, Ekaterina Kochmar, Preslav Nakov, and Fajri Koto. 2025. KazMMLU: Evaluating language models on Kazakh, Russian, and regional knowledge of Kazakhstan. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14403–14416, Vienna, Austria. Association for Computational Linguistics.

Unicode Consortium. 2025. Unicode security mechanisms. Unicode® Technical Standard #39, Version 17.0.0, eds. Mark Davis and Michel Suignard, `https://www.unicode.org/reports/tr39/`, file `confusables.txt`.

Maoli Wang, Xiaodong Zang, Jianbo Cao, Bowen Zhang, and Shengbao Li. 2024. PhishHunter: Detecting camouflaged IDN-based phishing attacks via siamese neural network. *Computers & Security*, 138:103668.

Lisa Yankovskaya, Maali Tars, Andre Tättar, and Mark Fishel. 2023. Machine translation for low-resource Finno-Ugric languages. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 762–771, Tórshavn, Faroe Islands. University of Tartu Library.

Ayur Zhanaev and Wojciech Połeć. 2019. A minority language in the globalizing world: The Buryat language on the Internet. *Adeptus*, 2019(14):1–17. Slavistics Institute of the Polish Academy of Sciences.

## A. *Palochka* Variant Filter Recall

To estimate the recall of the *palochka* variant filter for DCAD-2000 data discussed in Section 4 and to complement the precision estimates we modify our pipeline as follows: after converting DCAD-2000 documents to paragraphs, the first pipeline steps filter out short paragraphs, apply LID using the GlotLID model, and filter out all paragraphs in non-Cyrillic scripts. The paragraphs belonging to the four exception languages (Belarusian, Kazakh, Rusyn and Ukrainian) are then filtered out followed by the application of the *palochka* variant filter as a last step.

The paragraphs retained and discarded by the last filter are used to compute the overall and per-language values of filter recall as shown in Table 8. The total number of retained paragraphs (marked as true positives, or "TP") approximately corresponds to 1,628,526 paragraphs in NEC and NWC languages with *palochka* variants from Section 4. These numbers are not the same because

| Lang. | TP | FN | Recall (%) |
|---|---|---|---|
| abk | 6063 | 6 518 490 | 0.09 |
| abq | 14 329 | 6711 | 68.1 |
| ady | 136 171 | 24 137 | 84.94 |
| agx | 8009 | 95 943 | 7.7 |
| ava | 248 267 | 95 089 | 72.31 |
| che | 571 565 | 6 366 306 | 8.24 |
| dar | 21 025 | 15 654 | 57.32 |
| inh | 77 126 | 6 475 036 | 1.18 |
| kap | 539 | 1894 | 22.15 |
| kbd | 396 445 | 462 904 | 46.13 |
| lbe | 46 904 | 126 173 | 27.1 |
| lez | 72 608 | 198 008 | 26.83 |
| tab | 33 448 | 57 272 | 36.87 |
| tkr | 934 | 3847 | 19.54 |
| All | 1 633 433 | 20 447 464 | 7.4 |

Table 8: Numbers of paragraphs with (true positives, TP) and without (false negatives, FN) *palochka* variants and the corresponding value of recall.

the sequence of filters used to compute the recall is now different. Overall there are 20,447,464 paragraphs in relevant (according to GlotLID) languages without *palochka* variants shown as false negatives in the "FN" column of Table 8. The total value of recall is therefore 7.4%, which is significantly lower than the precision value of 17.3% computed in Section 4.

Several languages contribute to low recall. Abkhaz has negligible recall because *palochka* is not part of its orthography. Ingush and Chechen, along with Abkhaz, are the higher-resource languages among the 14 languages of interest. The low recall for Chechen and Ingush is likely due to a significantly more stable orthography and corresponding input methods, where the orthographically correct usage of *palochka* is preferred to its different variants. Finally, the relatively low recall for Aghul does not have a clear interpretation and further investigation is needed. The rest of the languages, including the lowest-resource Bezhta and Tsakhur, have reasonably high values of recall, where the best-performing languages are Abaza, Adyghe, Avar, Dargwa, and Kabardinian with a recall over 45%.

## B. LID Model-centric Experiments: An Outline

Wide-coverage LID models are trained to perform well across a range of languages, but often perform poorly on under-resourced languages which have a weak prior of appearing in the wild (*e.g.*, in web text). Can we use the presence of a *palochka* confusable, the *palochka* variant filter, as a high precision signal to *repair* predictions made by a modern LID model?

To investigate this question, the following experiment can be designed: take the dataset used to train the latest GlotLID v3.1 model[15] and partition it into a held-out set, where 15% of sentences are randomly held out per language, up to a maximum of 1,000 sentences per language.[16] For the training set, randomly sample up to 10,000 examples for each language from the remainder. Goot (2025) has observed that 1,000 sentences is sufficient for strong LID performance for most languages. Retrain a GlotLID fastText model, with hyperparameters given in (Kargaran et al., 2023), on this training set and generate the predictions on the held-out set.

After inference on the held-out set, collect all examples which were: (1) predicted to be a Cyrillic orthography outside of Belarusian, Ukrainian, Kazakh, and Rusyn;[17] (2) outside the set of NWC/NEC languages covered by GlotLID; and (3) passed the length constraint (contain at least three tokens). Note those that match the *palochka* variant filter, and count how many of these examples are actually labeled as a ground truth NWC/NEC language. The resulting proportion represents the *palochka* precision for North Caucasian orthographies. This experiment can be repeated multiple times to obtain a robust statistical estimate of the significance of computed precision values.

Another possible direction is to investigate the features of the GlotLID model itself, which takes hashed character $n$-gram features as input. If a particular orthographic feature, such as *palochka* letter, is useful according to the model, it will be included among the highly weighted features of GlotLID for NEC and NWC languages, either in a frequent unigram, or more likely a (hashed) character $n$-gram. The absence of such orthographic features that are hypothesized to be *a priori* informative from the list of top-weighted model features potentially indicates issues with the training data for the languages in question.

[15] https://huggingface.co/datasets/cis-lmu/glotlid-corpus

[16] As was described in the GlotLID paper.

[17] All of these employed a *palochka* confusable in their standard orthographies.