Research Journal DARYAFT
ISSN (E): 2616-6038, ISSN (P): 1814-2885
Volume 18. Issue 01, June 2026

تحقیقی وتنقیدی مجلّہ
دریافت

Received: 3rd February, 2026 | Accepted: 12th June, 2026 | Available Online: 30th June, 2026
Digital Object Identifier: 10.52015/daryaft.v18i01.441

پلیجیریزم ڈیٹکشن سافٹ ویئر، او سی آر اور اردو تحقیق: مسائل اور امکانات

Plagiarism Detection Software, OCR, and Urdu Research: Challenges and Possibilities

DR. ZAHOOR AHMAD
Secondary School Educator (Urdu), School Education Department, Punjab, Pakistan
(zahoorartist@gmail.com)

CONFLICT OF INTEREST: The author declares that there are no conflicts of interest related to the research, authorship, and/or publication of this article, and that the data presented have not been fabricated or falsified.

FUNDING: This research did not receive any specific grant or financial support from public, commercial, or not-for profit funding agencies.

PARTICIPANT CONSENT: The author confirms that Informed consent was obtained from all participants, and confidentiality was duly maintained.

KEYWORDS: Urdu Linguistics, OCR, Plagiarism, Urdu Corpus, Ligatures, Khat-e-Nastaliq

ABSTRACT: This article explores the relationship between technology and Urdu literary research, focusing on the challenges posed by plagiarism detection systems and Optical Character Recognition (OCR). Unlike English, a relatively ligature-free language, Urdu's cursive script and complex ligatures create significant difficulties for OCR development. At present, the absence of a comprehensive Urdu corpus allows a degree of flexibility in plagiarism detection, as a large body of classical and handwritten (calligraphic) Urdu material is not yet available in editable digital formats. The study classifies existing PDF formats of Urdu texts and evaluates the limitations of current OCR tools, including vFlat, Dastaan, and OCR developed by the Center for Language Engineering (CLE), particularly in handling diverse fonts and traditional calligraphy (Khat-e-Nastaliq). The development of a universal Urdu OCR is essential for building a robust Urdu corpus. Although this would increase scrutiny through plagiarism detection software, it would ultimately enhance academic standards in Urdu research by encouraging originality, critical engagement, and reduced reliance on unverified textual reproduction.

آج کل اردو لسانیات و ادبیات کے مقالے ایم ایس ورڈ میں لکھے جاتے ہیں۔ Plagiarism کے سوفٹ ویئر سے متعلقہ فائل کو گزارا جاتا ہے۔ اس سوفٹ ویئر سے پتہ چلتا ہے کہ محقق نے کتنے فی صد نقل و چسپاں (کاپی اینڈ پیسٹ) کیا ہے۔ اس بنا پر مقالے کے کامیاب یا ناکام ہونے کا فیصلہ کیا جاتا ہے یا ممتحن اس میں ترامیم کراتا ہے۔ ابھی تک اردو کے محققین عافیت میں ہیں۔ اردو کا سارا مواد

قابلِ ادارت(editable)صورت میں انٹرنیٹ پر شائع نہیں ہوا۔اس لیے انگریزی کے طلبہ کی نسبت اردو کے طلبہ کے لیے ابھی تک بہت سی آسانیاں موجود ہیں۔یہاں بصری حرف شناس(او سی آر)کا تذکرہ اہم ہے۔پہلے اس کی تعریف ملاحظہ ہو:

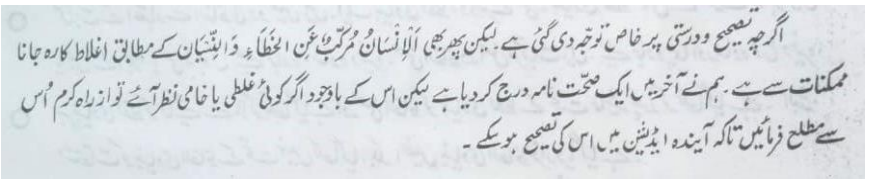
"Optical Character Recognition(OCR)وہ عمل ہے جو متن کی تصویر کو مشین کے لیے قابلِ قرات متن کے فارمیٹ میں تبدیل کرتا ہے۔مثال کے طور پر،اگر آپ کسی فارم یا رسید کو اسکین کرتے ہیں،تو آپ کا کمپیوٹر اسکین کو ایک تصویر فائل کے طور پر محفوظ کرتا ہے۔آپ تصویر فائل کی تدوین کرنے،تلاش کرنے،یا اس میں موجود الفاظ کو گننے کے لیے کسی ٹیکسٹ ایڈیٹر کا استعمال نہیں کر سکتے۔تاہم،آپ OCR کا استعمال کر کے تصویر کو ایک ٹیکسٹ ڈاکیومنٹ میں تبدیل کر سکتے ہیں جس کا مواد ٹیکسٹ ڈیٹا کے طور پر محفوظ ہو۔"(۱)

سادہ الفاظ میں بصری حرف شناس وہ سوفٹ ویئر ہے جو کسی متن کے عکس کو قابلِ ادارت متن میں بدلتا ہے۔انگریزی کی سب سے بڑی خوبی یہ ہے کہ یہ لگیچر فری زبان ہے۔یعنی انگریزی کے جو الفاظ لکھے جاتے ہیں،ان کے تمام حروف اپنی اشکال بر قرار رکھتے ہیں۔اس لیے بصری حرف شناس(او سی آر)کی مدد سے انگریزی الفاظ و حروف کے عکوس کو آسانی سے قابلِ ادارت متن میں بدلا جا سکتا ہے۔جو انگریزی کے بصری حرف شناس بنائے گئے ان میں تکنیکی پیچیدگیاں کم تھیں۔اس لیے ان کے ہاں کوئی مسئلہ نہیں تھا۔اردو میں سب سے بڑا مسئلہ اس کا رسم الخط ہے۔اس میں جب دو حروف آپس میں ملتے ہیں تو ان کی اشکال بدلنے کے امکانات موجود ہوتے ہیں۔اسے لگیچر کہتے ہیں۔ایک ویب کے مطابق:

> "گرافک ڈیزائن میں لگیچر(ترسیمہ)دو یا زیادہ حروف کو ایک ہی گلِف یا ٹائپو گرافک یونٹ میں ملانے کو کہتے ہیں۔"(۲)

مگر اردو میں بدلی ہوئی اشکالِ حروف سے بنا لفظ یا لفظ کا حصہ لگیچر یا ترسیمہ کہلاتا ہے۔مثال کے طور پر اردو کے دو حروف "ب" اور "د" کو ملایا جاتا ہے تو دونوں حروف سے بننے والے لفظ "بد" میں ب اور د کی شکلیں بدل جاتی ہیں۔فونٹ کی زبان میں اسے لگیچر یا ترسیمہ کہتے ہیں۔اردو ترسیمہ ہی بصری حرف شناس(او سی آر)کے لیے مسئلہ بنا۔اگرچہ اردو دورِ جدید سے ہم آہنگ ہو چکی ہے اور بہت سے فونٹس وجود میں آ چکے ہیں۔لیکن سارا مواد ان فونٹس میں کمپوز نہیں کیا گیا۔اردو کا زیادہ ذخیرہ کتابت اور خطاطی(ہاتھ سے کیا گیا کام)صورت میں موجود ہے۔مثال کے طور ایک ناول "فسانۂ عجائب" خطِ نستعلیق دہلوی میں ہے۔فیروز اللغات کی کتابت نستعلیق لاہوری میں کی گئی ہے۔کئی کتب ایسی ہو سکتی ہیں جو فارسی نستعلیق میں لکھی گئی ہیں۔اس لیے سوال پیدا ہوتا ہے کہ وہ بصری حرف شناس جو صرف فونٹس کے عکوس کو قابلِ ادارت متن میں بدل سکتا ہے،وہ کیسے کتابت کو قابلِ ادارت بنائے گا؟بصری حرف شناس جو اس وقت تک اہلِ اردو نے بنائے ہیں وہ مخصوص فونٹس پر ہی کام کر سکتے ہیں۔پس وہ محدود فوائد کے حامل ہیں۔ vFlat جو اہلِ مغرب کا ایجاد کردہ بصری حرف شناس ہے(۳)۔قدیم کتابت کے حامل نسخوں پر بھی کام کرتا ہے۔مگر اس میں مسئلہ یہ ہے کہ یہ سو صفحات کو ایک دن میں قابلِ ادارت متن میں بدلتا ہے جب کہ اس میں کتابت کے نسخوں پر کام کرتے ہوئے غلطی کے امکانات بھی موجود ہوتے ہیں۔البتہ تمام فونٹس پر یہ بہترین کام کر سکتا ہے۔اس کے محدودات سے انکار نہیں کیا جا سکتا۔

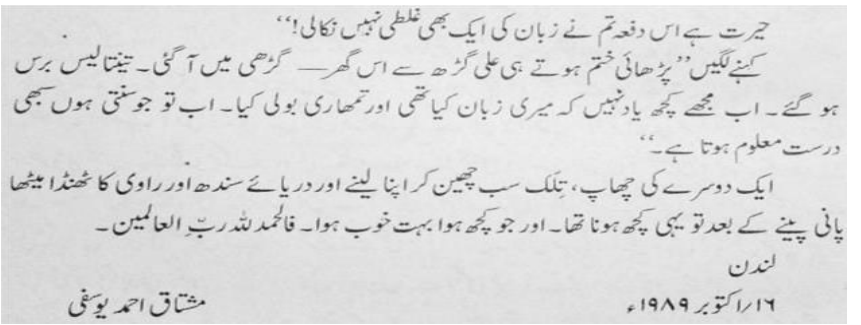
Research Journal DARYAFT
ISSN (E): 2616-6038, ISSN (P): 1814-2885
Volume 18. Issue 01, June 2026


اگرچہ تصحیح و درستی پر خاص توجہ دی گئی ہے۔ لیکن پھر بھی اَلْاِنْسَانُ مُرَکَّبٌ عَنِ الْخَطَاءِ وَالنِّسْیَان کے مطابق اغلاط کا رہ جانا ممکنات سے ہے۔ ہم نے آخر میں ایک صحّت نامہ درج کر دیا ہے لیکن اس کے باوجود اگر کوئی غلطی یا خامی نظر آئے تو ازراہ کرم اُس سے مطلع فرمائیں تاکہ آیندہ ایڈیشن میں اس کی تصحیح ہو سکے۔

(۱۔فیروز اللغات سے لیے صفحے کا عکس)

"اگرچہ تصحیح و درستی پر خاص توجہ دی گئی ہے۔ لیکن پھر بھی الانسان مُرکب عن الخطاء والنسیان کے مطابق اغلاط کا رہ جانا ممکنات سے ہے۔ ہم نے آخر میں ایک مصحتت نامہ درج کر دیا ہے لیکن اس کے باوجود اگر کوئی غلطی یا خامی نظر آئے تو ازراہ کرم اس سے مطلع فرمائیں تاکہ آیندہ ایڈیشن میں اس کی تصیح ہو سکے۔"

یہ وی فلیٹ سے کیا گیا تجربہ ہے۔ یہ عبارت کتابت کی صورت میں ہے۔ کتابت ہاتھ سے کی جاتی ہے۔ اگرچہ وی فلیٹ فونٹس کی پہچان کر لیتا ہے اور انھیں قابلِ ادارت متن میں تبدیل کر لیتا ہے۔ مگر جب اسے کتابت دی جائے تو اس میں غلطی کا امکان موجود رہتا ہے۔ مثال کے طور پر مذکورہ عبارت میں "صحت" کو مصحتت لکھ دیا گیا ہے۔ "اس" کو ماس اور "آئندہ" کو آئیندہ لکھا گیا ہے۔ تصحیح کو بھی غلط لکھا گیا ہے۔ یہ سوفٹ ویئر ایک لحاظ سے اچھا ہے کہ اس میں اشتہارات کم آتے ہیں۔ اس لیے کام روانی سے کیا جا سکتا ہے۔ مگر اس کا ایک مسئلہ ہے کہ یہ محدود صفحات کو یونی کوڈ میں بدلتا ہے۔ ایک ہی وقت میں ضخیم کتاب کو قابلِ ادارت متن میں نہیں بدلا جا سکتا۔ اس لیے یہ بھی وسیع کام کرنے کے لیے کامیاب تصور نہیں کیا جا سکتا۔ اوپر والے عکس کو متن میں وی فلیٹ میں تبدیل کیا ہے۔ یہ راقم کا تبدیل کردہ نہیں ہے۔ پاکستان میں ایک بصری حرف شناس بنایا گیا ہے۔ جس کے بانی سرمد حسین ہیں۔ یہ صرف نوری نستعلیق فونٹ کے ۱۴ تا ۴۴ پوائنٹ سائز کو پہچان سکتا ہے۔[۴] اس کا صرف نوری نستعلیق کے فونٹ کو قابلِ ادارت متن میں بدلنا اور پھر اسی فونٹ کے مخصوص سائز کو قابلِ ادارت متن میں بدلنا اس کے محدودات میں اضافہ کرتا ہے۔ یعنی اس کو وسیع پیمانے پر استعمال نہیں کیا جا سکتا اور نہ اس سے زیادہ فوائد اٹھائے جا سکتے ہیں۔ انجمن ترقیٔ اردو نے ایک بصری حرف شناس 'داستان' بنایا ہے۔[۵] مگر اس میں حروف کی پہچان کا عمل کامل نہیں ہے۔


حیرت ہے اس دفعہ تم نے زبان کی ایک بھی غلطی نہیں نکالی!''
کہنے لگیں ''پڑھائی ختم ہوتے ہی علی گڑھ سے اس گھر — گڑھی میں آ گئی۔ تینتالیس برس ہو گئے۔ اب مجھے کچھ یاد نہیں کہ میری زبان کیا تھی اور تمھاری بولی کیا۔ اب تو جو سنتی ہوں سبھی درست معلوم ہوتا ہے۔''
ایک دوسرے کی چھاپ، تِلک سب چھین کر اپنا لینے اور دریائے سندھ اور راوی کا ٹھنڈا میٹھا پانی پینے کے بعد تو یہی کچھ ہونا تھا۔ اور جو کچھ ہوا بہت خوب ہوا۔ فالحمد للہ ربّ العالمین۔
لندن
۱۶ر اکتوبر ۱۹۸۹ء
مشتاق احمد یوسفی

(۲۔مشتاق احمد یوسفی کی کتاب آبِ گم کے صفحے کا عکس)

حیرت ہے اس دفعہ تم نے زبانکی ایک بھی غلطی نہیں نکالی۔ کہنے لگیں" پڑھائی ختم ہوتے ہی علیگڑھ سے اسگھر گڑھی میں آگئی۔ شیتالیسبرس ہوگئے۔ اب مجھے کچھ یاد نہیں کہ میری زبان کیا تھی اور تمھاری بولی کیا۔ کہ اب توجوشیہوں سبھی درست معلوم ہوتاہے۔ ایک دوسرے کیچھپ، تکِسب چھین کر اپنالینے اور دریائے سندھ اور راوی کا ٹھنڈا میٹھ پانی پینے کے بعد تو یہی کچھ ہوناتھا۔ اور جو کچھ ہوا بہت خوب ہو۔ اسالمد اللہ ربِ نمین

لندن

17 اکتوبر 19ء

مشتا قاحمد یونھی

یہاں دیکھا جا سکتا ہے کہ اس عکس کو قابلِ ادارت متن میں بدلتے ہوئے کئی غلطیاں سامنے آئی ہیں۔ "زبان کی" کی جگہ "زبانکی"؛ سینتالیس کی جگہ "شیتالیس"؛ چھاپ تلک سب کی جگہ "چھیپ تلکسب" اور "الحمد للہ رب العالمین" کی جگہ "اسالمد اللہ ربانمین" لکھا ہے۔ مشتاق احمد یوسفی کا نام بھی غلط ہو گیا ہے۔ لہذا اس بصری حرف شناس سے فی الوقت فائدہ اٹھانا ممکن نہیں ہے۔ اس کے محدودات کو نظر انداز نہیں کیا جا سکتا۔

OCRx[۲] کی مدد سے مذکورہ عکس کو قابلِ ادارت متن میں یوں بدلا گیا: (یہ راقم کا لکھا گیا متن نہیں ہے۔ یہ OCRx کا تبدیل کردہ متن ہے)

حیرت ہے تم نے اس دفعہ زبان کی ایک غلطی بھی نہیں نکالی! کہنے لگی: "پڑھائی ختم ہوتے ہی علی گڑھ سے گھر گڑھی میں آگئی۔ تینتالیس برس ہوگئے۔ اب مجھے کچھ یاد نہیں کہ میری زبان کیا تھی اور تمہاری بولی کیا۔ اب تو جو سنتی ہوں کبھی درست معلوم ہوتا ہے۔ "ایک دوسرے کی چھاپ تلک سب چھین کر اپنا لینے اور دریائے سندھ اور راوی کا میٹھا پانی پینے کے بعد تو یہی کچھ ہونا تھا۔ اور جو کچھ ہوا بہت خوب ہوا۔ فالحمد للہ رب العالمین۔

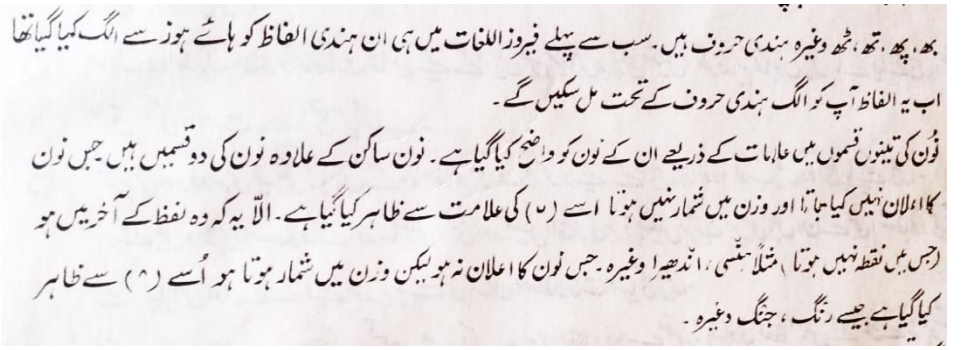
لندن

۱۶ اکتوبر ۱۹۸۹ء

مشتاق احمد یوسفی

اس بصری حرف شناس کا نتیجہ سو فی صد ہے۔ لیکن یہ سوفٹ ویئر اردو کتابت کو قابلِ ادارت متن میں نہیں بدل سکتا۔ اس لیے اردو کے قدیم سرمائے کو اس کی مدد سے کوئی فائدہ نہیں پہنچ سکتا۔ اگر فونٹ والے نکوس کو بھی قابلِ ادارت بنایا جائے تو وقت زیادہ لگ سکتا ہے۔ اور اس کا ضمنی اثر یہ بھی ہے کہ محدود صفحات کی پہچان کر سکتا ہے۔ سیکڑوں اور ہزاروں صفحات پر مبنی کتاب کو ایک ہی وقت میں یونی کوڈ میں نہیں لایا جا سکتا۔

Research Journal DARYAFT
ISSN (E): 2616-6038, ISSN (P): 1814-2885
Volume 18. Issue 01, June 2026



بھ، پھ، تھ، ٹھ وغیرہ ہندی حروف ہیں۔ سب سے پہلے فیروز اللغات میں ہی ان ہندی الفاظ کو ہائے ہوز سے الگ کیا گیا تھا اب یہ الفاظ آپ کو الگ ہندی حروف کے تحت مل سکیں گے۔

نُون کی تینوں قسموں میں علامات کے ذریعے ان کے نون کو واضح کیا گیا ہے۔ نون ساکن کے علاوہ نون کی دو قسمیں ہیں جس نون کا اعلان نہیں کیا جاتا اور وزن میں شمار نہیں ہوتا اسے (ہ) کی علامت سے ظاہر کیا گیا ہے۔ الّا یہ کہ وہ لفظ کے آخر میں ہو (جس میں نقطہ نہیں ہوتا) مثلاً ہنسی، اندھیرا وغیرہ۔ جس نون کا اعلان نہ ہو لیکن وزن میں شمار ہوتا ہو اُسے (ں) سے ظاہر کیا گیا ہے جیسے رنگ، جنگ وغیرہ۔

(۳۔فیروز اللغات کے صفحے کا ایک عکس)

بھ چھ، تھ، ٹھ وغیرہ ہندی حروف ہیں۔ سب سے پہلے فیروز اللغات میں ہی ان مہندی الفاظ کو ہائے ہوز سے الگ کیا گیا تھا اب یہ الفاظ الفاظ آپ کو الگ ہندی حروف کے تحت مل سکیں گے۔ کرافون کی تینوں قسموں میں علامات کے ذریعے ان کے نون کو واضح کیا گیا ہے۔ نون ساکن کے علاوہ نون کی دو قسمیں ہیں جس نون کا اعلان نہیں کیا جاتا اور وزن میں شمار نہیں ہوتا اسے (ہ) کی علامت سے ظاہر کیا گیا ہے۔ الا یہ کہ وہ لفظ کے آخر میں ہو جس میں نقطہ نہیں ہوتا، مثلاً سنسی، اندھیرا وغیرہ جس نوں کا اعلان نہ ہو لیکن وزن میں شمار ہوتا ہو اُسے (ں) سے ظاہر کیا گیا ہے جیسے رنگ، جنگ وغیرہ۔

گوگل لینز[۷] کو اوپر والا عکس دیا گیا۔ (یہ بھی راقم نے نہیں لکھا۔ یہ گوگل لینز کا تیار کردہ متن ہے) اس کا نتیجہ دیکھا جا سکتا ہے۔ اس نے کتابت کی پہچان خوب کی ہے۔ لیکن ہندی کو مہندی لکھا گیا ہے۔ الفاظ کو دو بار لکھا گیا ہے۔ نون کی جگہ فون لکھ دیا گیا ہے۔ ہنسی کو سنسی لکھا گیا ہے۔ نون کو نوں لکھا گیا ہے۔ پہچان کی خاطر غلط الفاظ کو دبیز (بولڈ) کر دیا گیا ہے۔ ن غنہ اور ن کے لیے وضع کی علامت کو اس سوفٹ ویئر نے اردو ہندسوں میں بدل دیا ہے۔ اس کو اگر کسی فونٹ والی عبارت پر آزمایا جائے تو اس کے نتائج بہت بہتر آتے ہیں اور ان میں غلطیاں کم ہوتی ہیں۔ کتابت کے نمونے اس لیے ظاہر کیے جا رہے ہیں کیوں کہ اردو کا قدیم سرمایہ کتابت کی صورت میں موجود ہے۔ اور اس کی مقدار بھی زیادہ ہے۔ اس کو قابلِ ادارت متن میں بدلنے کی اشد ضرورت ہے۔

جب آن لائن گوگل لینز سے کسی عبارت کے عکس کو قابلِ ادارت متن میں بدلا جاتا ہے تو ایک ایک صفحے کو عکس میں ڈھال کر گوگل لینز پر اپ لوڈ کیا جاتا ہے اور اسے قابلِ ادارت متن بنایا جاتا ہے۔ اس کا منفی پہلو یہ ہے کہ اس میں وقت کا ضیاع ہوتا ہے۔ ۵۰۰ صفحات کی کتاب کو قابلِ ادارت متن میں لانے کے لیے کافی وقت لگ سکتا ہے۔ اس لیے اس کو ترجیح نہیں بنایا جا سکتا۔ چند صفحات کے مضمون اور کتاب کے لیے یہ مناسب ہے۔ اور یہ تمام قسم کے عکوس کو قابلِ ادارت متن نہیں بنا سکتا۔ سو اسے مثالی سوفٹ ویئر نہیں سمجھا جا سکتا۔ اس لیے مثالی سوفٹ ویئر وہی ہے جو کم از کم ایک ہزار صفحات تک کو ایک ہی وقت میں قابلِ ادارت متن میں بدل سکے۔ اس سے وقت کی بچت بھی ہو سکتی ہے اور زیادہ زیادہ سے کتب کو قابلِ ادارت متن کی صورت میں بدلا جا سکتا ہے۔ اس کے علاوہ آن لائن مارکیٹ میں سوفٹ ویئرز اطلاقیوں کی صورت میں بھی موجود ہیں۔ ان میں بھی محدودات پائے جاتے ہیں۔

اچھے بصری حرف شناس کم ہونے یا نہ ہونے کی وجہ سے اردو کے تمام متون کے عکوس کو قابلِ ادارت متن کی صورت میں انٹرنیٹ پر نہیں لایا جاسکا۔ پس اردو "کارپس" کی صورت میں ہمارے پاس موجود نہیں ہے۔ کارپس کی تعریف یہ ہے: "کارپس دراصل زبان کے متعلق ٹیکنالوجی سے متعلق ایسا ذخیرہ ہے جو کمپیوٹر پر استعمال ہونے والے مواد پہ مشتمل ہوتا ہے۔"(۸) یہ مواد تحریری صورت (قابلِ ادارت)، صوتی صورت اور ویڈیو کی شکل میں بھی ہو سکتا ہے۔ لیکن قابلِ ادارت مواد کو تحقیق میں اہم سمجھا جاتا ہے۔

پاکستان میں اردو کارپس بنانے کی ایک کوشش ماضی میں سامنے آچکی ہے۔ ادارہ فروغ قومی زبان، جس کا قدیم نام مقتدرہ قومی زبان تھا۔ اس وقت کے پروگرامر رضوان عزیز نے "اردو کوائفیہ" بنایا تھا۔ اس میں ایک املا کے الفاظ کی فہرست سامنے آتی تھی۔ اس میں اعراب کی مدد سے مختلف معانی اور مفاہیم کو واضح کیا جاتا تھا۔ تجنیسِ حرفی کے ذخائر بھی موجود تھے۔ صارف کو جس حرف کی ضرورت ہوتی تھی، وہاں سے مطلوبہ الفاظ کا چناؤ کیا جاسکتا تھا۔ تعددِ امثال کی مدد سے کلیدی الفاظ، دیگر املا، تذکیر و تانیث، واحد جمع اور ہم ردیف کی فہرست آتی تھی۔ اس کے علاوہ کتاب اور شاعر کا زمرہ بھی موجود تھا۔ یعنی کسی شاعر کے بارے میں معلومات حاصل کرنا آسان تھا۔ صارف اپنا خیال پیش کر سکتا تھا۔ مشین اس کا تجزیہ کرتے ہوئے سارے جملے کو اپنے مقرر کردہ فارمولے کے مطابق پروسیس کر سکتی تھی اور اس کے ممکنہ نتائج تعدد کے مطابق پیش کر سکتی تھی۔(۹) لیکن یہ بڑا منصوبہ حکومتی عدم توجہی کا شکار ہونے کی وجہ سے منظرِ عام پہ نہ آسکا۔(۱۰)

اردو کارپس بن جائے تو اردو کا زیادہ ذخیرہ قابلِ ادارت مواد تحریری صورت میں انٹرنیٹ پر موجود ہو گا۔ یعنی اسے نقل و چسپاں کیا جاسکے گا۔ اور یہ پلیجرزم (سرقہ نویسی) کے سوفٹ ویئرز کی پہنچ میں ہو گا۔

اگرچہ محققین کے لیے یہ آسانی تو رہتی ہے کہ ان کی پلیجرزم کی فی صدی HEC کے معیار کے مطابق درست ہوتی ہے۔ اسے تحقیق کے اعلیٰ معیار کے مطابق دیکھا جائے تو بذاتِ خود یہ ایک نقص ہے جو تحقیق کے معیار کو کم زور بنا دیتا ہے۔ موجودہ دور میں اردو کی اکثر کتب پی ڈی ایف میں بدل کر انٹرنیٹ پر آچکی ہیں۔ محققین کے لیے اس ضمن میں وہ آسانی سے اپنے متعلقہ مواد کو ان کتابوں سے لے لیتے ہیں یوں انھیں کتب خانوں میں جانے کی زیادہ ضرورت نہیں پڑتی اور وقت اور رقم کی بچت بھی ہو جاتی ہے۔ جامعات محققین کو تحقیقی مقالہ مقالہ مکمل کرنے کے لیے ایک سال (ایم فل) یا دو سال (پی ایچ ڈی) کا وقت دیتی ہیں۔ اس لیے اگر موضوع تکنیکی قسم کا ہو یا پیچیدہ، تو مقالے کی تکمیل مشکل ہو جاتی ہے۔ لہٰذا محققین آج کل آن لائن کتابوں پر انحصار کرتے ہیں، جن میں سے اکثر پی ڈی ایف میں ہوتی ہیں۔ پی ڈی ایف کی تعریف ملاحظہ ہو:

"پی ڈی ایف ایک ایسا فارمیٹ ہے جو پرنٹ شدہ ڈاکومنٹ کے تمام اجزا کو اپنی گرفت میں لاتا ہے اور انھیں عکس کی شکل دیتا ہے۔ جسے دیکھا جاسکتا ہے، پرنٹ کیا اور دوسروں کو بھیجا جاسکتا ہے۔"(۱۱) اس کے بہت سے فوائد ہیں۔ اس کی مدد سے کسی بھی فائل کو آسانی سے پڑھا جاسکتا ہے۔ پی ڈی ایف کی تین صورتیں ہو سکتی ہیں:

اول: وہ عبارت جو ان پیج کے پرانے ورژن میں لکھی گئی اور اسے پی ڈی ایف میں بدلا گیا۔

دوم: وہ عبارت جس کے عکس لے کر انھیں پی ڈی ایف کی شکل دی گئی۔

سوم: وہ متن جو ایم ایس ورڈ اور ان پیج کے نئے ورژن میں لکھا گیا اور اس کو پی ڈی ایف میں تبدیل کیا گیا۔

پہلی صورت کا مسئلہ یہ ہے کہ اسے ان پیج سے ایم ایس ورڈ میں لایا جائے تو یہ عجیب و غریب اشکال سامنے لاتا ہے۔ اس کی بنیادی وجہ یہ ہے کہ پرانا ان پیج یونی کوڈ میں نہیں تھا۔ یونی کوڈ کی تعریف درج ذیل ہے:

> "یہ ایک بین الاقوامی انکوڈنگ معیار ہے جو مختلف زبانوں اور رسم الخط کے ساتھ استعمال کے لیے ہے، جس کے ذریعہ ہر حرف، عدد، یا علامت کو ایک منفرد عددی قیمت دی جاتی ہے جو مختلف پلیٹ فارمز اور پروگراموں میں لاگو ہوتی ہے۔"(۱۲)

یونی کوڈ ایسا نظام ہے جو زبانوں کو عالمی سطح پر کمپیوٹر کو سمجھنے میں مدد دیتا ہے۔ اس لیے جو زبان یونی کوڈ کے نظام میں آجاتی ہے، اسے کمپیوٹر پر دنیا کے کسی بھی کونے میں پڑھا جا سکتا ہے اور کمپیوٹر اسے سمجھتا ہے۔ چونکہ اردو یونی کوڈ میں ہے اس لیے کمپیوٹر یا انٹرنیٹ کی زبان بن چکی ہے۔ جب ان پیج کے پرانے ورژن موجود تھے تو ان میں یونی کوڈ نہیں تھا۔ اس لیے ان کو ان پیج سے کاپی کیا جاتا تھا تو کمپیوٹر انھیں نہیں سمجھتا تھا۔

دوسری صورت میں تو عکس کو پی ڈی ایف میں بدلا گیا ہے، اس کو قابلِ ادارت متن میں بدلنے کا امکان ہی نہیں ہے۔

تیسری صورت میں اس بات کا امکان ہے کہ پی ڈی ایف کو نقل کر کے ان پیج جدید یا ایم ایس ورڈ میں چسپاں کیا جائے تو وہ عبارت قابل ادارت صورت میں سامنے آسکتی ہے۔ مگر یہ امکان صرف کمپیوٹر پر موجود ہے۔ اینڈرائیڈ سے ایسا نہیں ہو سکتا۔ مثال کے طور پر یہ عکس اس عبارت کا ہے جسے یونی کوڈ ورژن سے پی ڈی ایف میں بدلا گیا ہے:

کمپیوٹر اور انٹرنیٹ پر اُردو

Ali Sheraz

عصر حاضر میں کمپیوٹر اور انٹرنیٹ ایک بنیادی ضرورت اور ہماری زندگی کا جزو لاینفک بن چکا ہے۔ خیالات کی ترسیل کے لئے پہلے خطوط لکھے جاتے تھے اور کاغذ انسانی ہاتھ کے لمس کی مہک سے معمور ہوتے تھے۔ کاغذ کے دور سے کی بورڈ، کتاب کے دور سے ای کتاب اور خط کے دور سے ای میل تک کا سفر انسان نے بہت جلد طے کر لیا۔ یہ سب کچھ جدید ٹیکنالوجی کمپیوٹر اور انٹرنیٹ کی بدولت ہوا۔ فاصلوں کا جھنجٹ ہی ختم ہوا اور جغرافیائی سرحدیں اپنا سا منہ لے کر رہ گئیں۔

اردو کمپیوٹنگ دراصل کمپیوٹر پر اردو استعمال کرنے اور انفارمیشن ٹیکنالوجی کے میدان میں اردو کے متعلق اور اردو میں تعلیم و تحقیق کا نام ہے۔ برصغیر میں ٹائپ رائٹر کی آمد کے تھوڑے عرصے بعد ہی اردو بھی ٹائپ رائٹر سے لکھی جانے لگی لیکن ٹائپ رائٹر سے صرف نسخ رسم الخط یعنی نسخ فانٹ میں ہی اردو لکھی جا سکتی تھی۔ بلکہ مولانا ابوالکلام آزاد کی کچھ تصانیف کی کتابت بھی ٹائپ رائٹر کے ذریعے ہوئی تھی۔

نستعلیق رسم الخط تکنیکی لحاظ سے تھوڑا پیچیدہ ہے کیونکہ اگر ہم مشین پر نستعلیق لکھنے کے حوالے سے بات کریں تو جیسے جیسے کسی لفظ کے میں حروف کا اضافہ ہوتا ہے ویسے ویسے پچھلے حروف نئے لکھے گئے حرف کے مطابق اپنی شکلیں اور جگہیں تبدیل کرتے ہیں۔ نستعلیق کی ایسی پیچیدگیوں کی وجہ سے ماضی میں کئی لوگوں نے یہاں تک کہا تھا کہ اردو کا معیاری رسم الخط فارسی والوں کی طرح نستعلیق سے نسخ کر دینا چاہئے۔

(۴۔ ذاتی کمپیوٹر پر ایم ایس آفس میں لکھا ہوئے مضمون کا عکس؛ جس کو پی ڈی ایف میں بدلا گیا۔)

Research Journal DARYAFT
ISSN (E): 2616-6038, ISSN (P): 1814-2885
Volume 18. Issue 01, June 2026

تحقیقی وتنقیدی مجلّہ
دریافت

ویپمکر ٹاورار ٹنٹینر پُاردو

Ali Sheraz

رصعاحرضہں یموپیمکر ٹاورار ٹنٹینا کیاینبدیر ضور تاورامہریز دنیگا کزجو الکفنیبا کچے ہ۔

ایخالتیکر تلیسے کے ئلے لہپو طخطے ھکلاجے تے ھتاوراکذغا اسینا ھتے کسملیکھے سو معمرو ہے تے ھت۔

اکذغے کدورے سیکوبرڈ، اتکبے کدورے سایاتکباور طحنے کدورے سایلیمکتا کرفسلا اسنمے نتہبد لجے طر کایل۔

ہیبسھچکد جدیے لرکرہں یمنگ۔ انکیٹو لیجو پیمکر ٹاورار ٹنٹینیکد بو تلوہا۔ افولصں ا کنجبھجیہمتخوہا اور رغجا ایفیئر سدحں یا انپاسُ مُہیپر ارٹئیکا آ

دمے کاردوو پیمکگنٹدر الصو پیمکر ٹرپاردواا معتسکر کے ناوراافنر نشیما نکیٹو لیجے کدیماں یمپاردوے کقلعتما وراردوں یمیملعتو قیقحتا کانلے ہ۔ر

بیغں یما ٹئیپر ارٹئے سر صفخسنر مسا ظحلینعیخسنا فٹنں یمیہاردو یھکلا جیتکسیھت۔ ہکلبو مالانا وبا الکلاّزادیکھچک

پپر ارٹئے سیھکلاجے نیگلنکیلا ٹئو ھتڑے رعے صدعبیہار دویھبا ٹئیپر ارٹئے کذرے عیوہیئ...

اس عبارت کو اینڈرائیڈ پر آزمایا گیا۔ اس طرح کی پی ڈی ایف عبارات کو اینڈرائیڈ سے قابلِ ادارت متن میں تبدیل کرنے کی کوشش کی جائے تو مندرجہ بالا صورتِ حال نظر آتی ہے۔ اس لیے یہ کام اینڈرائیڈ سے ممکن نہیں ہے۔

سو اس معاملے میں پی ڈی ایف محققین کی مدد نقل و چسپاں میں نہیں کر سکتا۔ نتیجتاً پی ڈی ایف موجودہ تجربات کی روشنی میں مدد نہیں کر سکتی۔ آخر میں بصری حرف شناس کا ہی انتخاب باقی رہ جاتا ہے۔

اگر اردو کے ماہرین یا تکنیکی سائنس کے لوگ اردو کا ایسا بصری حرف شناس بنا لیتے ہیں جو تمام فونٹس کے ساتھ ساتھ کتابت کے تمام اسالیب کی پہچان کر لیتا ہے تو اردو کارپس با آسانی بنایا جا سکتا ہے۔

اس کا اثر یہ ہو گا کہ ایک محقق نے جہاں جہاں سرقہ کیا ہو گا، اس کی تمام تفاصیل پلیجرزم کے سوفٹ ویئر میں آ جائیں گی۔ اس طرح ایک محقق کے لیے جہاں مسائل پیدا ہوں گے وہاں تحقیق کے معیار میں اضافہ ہو گا۔ ایک محقق جو محض نقل و چسپاں پر انحصار کرتا تھا، اسے تنقیدی بصیرت کو استعمال میں لانا ہو گا۔ لیکن ایک مسئلہ پھر درپیش ہو گا کہ تخلیقی یا اقتباسی تحریر اور سرقہ میں فرق کرنا مشکل ہو جائے گا۔ اس حوالے سے سوفٹ ویئرز کو تربیت کے لیے نمونے درکار ہوں گے، جن کی مدد سے وہ درست سمت میں اپنا کام کر سکے۔ کچھ سفارشات اور نتائج اس ضمن میں پیش کیے جاتے ہیں:

الف۔ اردو کے بصری حرف شناس پر اردو کے ماہرین اور تکنیکی علوم کے جاننے والے کام کر سکتے ہیں۔ اس طرح وہ ایسا بصری حرف شناس بنا سکتے ہیں جو کتابت کے تمام اسالیب اور ہر قسم کے اردو فونٹس کو قابلِ ادارت (editable) بنا سکے۔ اس کے علاوہ جو ویب سائیٹس پی ڈی ایف ورژن کی کتابیں اپ لوڈ کرتی ہیں، ان سے آئیڈیا لیا جا سکتا ہے۔ اس کی بڑی مثال ریختہ ہے۔ اس پر ہزاروں نایاب کتب اور جدید کتابیں پی ڈی ایف میں موجود ہیں۔ اس طرح کے بڑے منصوبوں میں کئی رضاکاروں کو شامل کیا جاتا ہے۔ جو پہلے کتابوں کو سکین کرتے ہیں اور بعد ازاں پی ڈی ایف کی شکل میں لاتے ہیں۔ اگر حکومت یا کوئی ادارہ ایسا منصوبہ ترتیب دیتے ہیں تو ایک گروہ کے سیکڑوں رضاکاروں کی مدد سے کئی کتابیں روزانہ کی بنیاد پر سکین (scan) ہو سکتی ہیں۔ دوسرا گروہ ان افراد پر مشتمل ہو جو ان کو

مختلف بصری حرف شناس استعمال کر کے ان کو قابلِ ادارت متن میں بدل سکتا ہے۔ تیسرا گروہ اغلاط کی درستی (پروف ریڈنگ) کر کے اسے حکومت یا ادارے کی متعلقہ ویب سائیٹ پہ اپ لوڈ کر سکتا ہے۔ تسلسل سے کام کیا جائے تو کتابوں کا ایک اچھا خاصا ذخیرہ بن سکتا ہے۔ اس طریقے سے بھی اردو کارپس بھی بن سکتا ہے۔

ب۔ اگر ہمہ جہت بصری حرف شناس بن جاتا ہے تو اردو کے محقق کو زیادہ سنجیدہ ہونا پڑے گا۔ ورنہ اس کا متن سرقے کے زمرے میں آجائے گا۔

ج۔ اردو میں اگر بصری حرف شناس آجاتا ہے جس میں تمام مسائل کا حل ہو تو اردو تحقیق کا معیار عالمی سطح کا ہو جائے گا۔

اس وقت مختلف ادارے اردو کارپس پر کام کر رہے ہیں۔ مثال کے طور پر منہاج یونیورسٹی لاہور نے اس سلسلے میں کام شروع کیا ہوا ہے۔ جسے "کارپس ریسرچ سنٹر" کا نام دیا گیا ہے۔ یہ اردو کارپس کے ساتھ ساتھ مقامی زبانوں کے کارپس پر بھی کام کر رہا ہے۔[۱۳] سنٹر فار لینگویج انجینئرنگ (Center for Language Engineering) بھی اس سلسلے میں کام کر رہا ہے۔[۱۴] اس کا کام کمپیوٹیشنل لسانیات (Computational Linguistics)، میڈیا کارپس، اردو اے ایس آر (Urdu ASR)، آواز کی پہچان (Speech recognition) وغیرہ پر جاری ہے۔ اس طرح پاکستان کے مشہور ادارے NUST میں ایک ایسے بصری حرف شناس پہ کام ہو رہا ہے جو ہاتھ کی لکھائی کو پہچان سکتا ہے۔ جسے UHWR-NUST کا نام دیا گیا ہے۔[۱۵] ایک اور بصری حرف شناس اردو بان کے نام سے انٹرنیٹ پر موجود ہے[۱۶] یہ اوپن سورس ہے اور مفت ملتا ہے۔ اس کی مدد سے نستعلیق اور نسخ کے عکوس کو یونی کوڈ میں بدل سکتا ہے۔ اس میں غلطیوں کے امکانات کم ہیں۔ اس طرح ادارے کام کرتے رہے تو اردو کارپس کا مستقبل خوش آیند ہو سکتا ہے اور اردو کے تحقیقی معیار میں خاطر خواہ اضافہ ہو سکتا ہے۔

حواشی وحوالہ جات

۱۔ Amazon Web Services. "What Is OCR?" https://aws.amazon.com/what-is/ocr/ تاریخ استفادہ ۶ جون ۲۰۲۵ء

۲۔ The Brief. "What Is Ligature?" https://www.thebrief.ai/blog/what-is-ligature/ تاریخ استفادہ ۲ فروری ۲۰۲۶ء

۳۔ SourceForge. "vFlat Scan." https://sourceforge.net/software/product/vFlat-Scan/# تاریخ استفادہ ۲ فروری ۲۰۲۶ء

۴۔ Center for Language Engineering. "OCR." https://www.cle.org.pk/ocr/ تاریخ استفادہ ۲ فروری ۲۰۲۶ء

۵۔ داستان، "اردو ہے میرا نام" https://dastaan.io/ تاریخ استفادہ ۲۴ جولائی ۲۰۲۵ء

۶۔ FocalStudio. "OCRx." https://github.com/FocalStudio/OCRx. تاریخ استفادہ ۲ جولائی ۲۰۲۳ء

۷۔ Google. "Google Lens." https://www.lens.google: (فیروز اللغات کا ایک صفحہ گوگل لینز کو پہچان کے لیے دیا گیا) ۲۷ مئی ۲۰۲۶ء

۸۔ اردو کارپس کی تیاری، متن کا تجزیہ اور غلطیوں کے نمونوں کی درجہ بندی، تاریخ استفادہ یکم دسمبر ۲۰۲۵ء ib.bazmeurdu.net.

۹۔ رضوان عزیز، اردو کو ائفیہ یا ڈیٹا بیس: اردو تحقیق کا بنیادی آلہ۔ "اخبارِ اردو، مقتدرہ قومی زبان، اسلام آباد، جولائی ۲۰۰۹ء

۱۰۔ مصاحبہ، رضوان عزیز (گرافک ڈیزائنر، ادارۂ فروغ قومی زبان) اسلام آباد، بوقت ۱۱ بجے دن، ۱۷ اپریل ۲۰۲۴ء

۱۱۔ TechTarget, Portable Document Format (PDF), https://www.techtarget.com/whatis/definition/Portable-Document-Format-PDF

۱۲۔ Authority Entry, Oxford Reference, https://www.oxfordreference.com/display/10.1093/oi/authority.20110803110655843 / تاریخ استفادہ ۱۲ اپریل ۲۰۲۱

۱۳۔ CRC Multan. www.crc.mul.edu.pk. ۲۷ مئی ۲۰۲۶ء

۱۴۔ KICS Lahore. www.kics.edu.pk ۲۷ مئی ۲۰۲۶ء

۱۵۔ A Unified Architecture for Urdu Printed and Handwritten Text Recognition, PDF Document. ۲۷ مئی ۲۰۲۶ء

۱۶۔ Urdubaan, Enterprise-Grade Urdu OCR and TTS ۲۷ مئی ۲۰۲۶ء

References in Roman Script:

1. Amazon Web Services, What Is OCR?, Amazon, https://aws.amazon.com/what-is/ocr/. Accessed 6 June 2025
2. The Brief, What Is Ligature?, The Brief Blog, 2 Feb. 2026, https://www.thebrief.ai/blog/what-is-ligature/
3. SourceForge, vFlat Scan, SourceForge, 2 Feb. 2026, https://sourceforge.net/software/product/vFlat-Scan/#.
4. Center for Language Engineering, OCR, CLE Pakistan, 3 Feb. 2025, https://www.cle.org.pk/ocr/.
5. Dastan:Urdu hey mera Naam, Dastaan, 24 July 2025, https://dastaan.io/
6. FocalStudio, OCRx, GitHub, 2 July 2023, https://github.com/FocalStudio/OCRx.
7. Google, Google Lens, Google, 27 May 2026, https://www.lens.google.
8. Bazm-e-Urdu, Preparation of the Urdu Corpus, Text Analysis, and Classification of Error Patterns, Bazm-e-Urdu, 1 Dec. 2025, ib.bazmeurdu.net/.
9. Aziz, Rizwan, Urdu Koāfiya or Database: A Fundamental Tool for Urdu Research,

Akhbar-e-Urdu, Muqtadra Qaumi Zaban, Islamabad, July 2009.
10. Aziz, Rizwan. Interview. Graphic Designer, Idara-e-Farogh Qaumi Zaban, Islamabad, 17 Apr. 2024, 11:00 a.m.
11. TechTarget, Portable Document Format (PDF), TechTarget, https://www.techtarget.com/whatis/definition/Portable-Document-Format-PDF.
12. Oxford Reference, Authority Entry, Oxford University Press, 2 Apr. 2021, https://www.oxfordreference.com/display/10.1093/oi/authority.20110803110655843/.
13. CRC, CRC Multan, CRC, 27 May 2026, www.crc.mul.edu.pk.
14. KICS, KICS Lahore. KICS, 27 May 2026, www.kics.edu.pk.
15. A Unified Architecture for Urdu Printed and Handwritten Text Recognition, PDF Document, 27 May 2026
16. Urdubaan, Enterprise-Grade Urdu OCR and TTS, Urdubaan, 27 May 2026

Dr. Zahoor Ahmad is a Secondary School Educator (Urdu) at the School Education Department, Punjab, Pakistan. He earned his PhD from Qurtuba University of Science and Information Technology Peshawar. His academic specialization is Urdu Linguistics, and he has one published research article to his credit.