कम-संसाधनीय भाषाओं के ओसीआर का एक संक्षिप्त सर्वेक्षण
(A Concise Survey of OCR for Low-Resource Languages)

मिलिंद अगर्वाल **अंतोनिओस अनस्तासोपोलोस**
Milind Agarwal Antonios Anastasopoulos
कंप्यूटर विज्ञान विभाग (Department of Computer Science)
जॉर्ज मेसन विश्वविद्यालय (George Mason University)
{magarwa,antonis}@gmu.edu

सारांश (Abstract)

आधुनिक प्राकृतिक भाषा प्रसंस्करण (एनएलपी) तकनीकों को समय के साथ साथ ठोस कलनविधियों (एल्गोरिदम) को प्रशिक्षित करने के लिए सारवान मात्रा में डेटा की आवश्यकता पड़ती जा रही है। कम-संसाधनीय भाषाओं में ऐसी प्रोद्योगिकियों का निर्माण करना तभी मुमकिन है जब डेटा सृजन प्रयास और डेटा-दक्ष (कार्यकुशल) कलनविधियों पर ध्यान केंद्रित किया जाये। बहुत सी कम-संसाधनीय भाषाओं के लिये, ख़ास तौर से महाअमेरिका की देशी भाष्यों के लिए, यह डेटा प्रतिबिंब-आधारित मशीन-अपठनीय दस्तावेज़ों में मौजूद है। उधारणतः, व्यापक शब्दकोशों की स्कैण्ड प्रतियाँ, भाषाशास्त्रीय फील्ड नोट्स, बालकथाएँ, और अन्य पाठ्य सामग्री। इन संसाधनों (डेटा) के डिजिटलीकरण के लिए, प्रकाशिक सम्प्रतीक अभिज्ञान (ओसीआर) ने एक मुख्य भूमिका निभाई है पर कम-संसाधनीय विन्यासों में इसके साथ कई मुश्किलें भी साथ आती हैं। इस पेपर में, हम इन विन्यासों के लिए विशिष्ट ओसीआर तकनीकों का पहला सर्वेक्षण साझा करते हैं, और महामेरिका की देशी भाषाओं पर विशिष्ट बल केंद्रित कर, कई खुली चुनौतियों की रूपरेखा बनाते हैं। हमारे पूर्व अनुसंधानों और नतीजों के आधार पर, हम अभिकलनात्मक शोधकर्ताओं, भाषावैज्ञानिकों, और भाषाई समुदायों के फ़ायदे के लिए ओसीआर के उपयोग और उसके सुधार के लिए कुछ सिफ़ारिशों के साथ, इस पेपर को निष्कर्षित करते हैं।

1 भूमिका

लातिनी अमेरिका सैकड़ों देशी भाषाओं के एक भाषाई रूप से विविध समुच्चय का गढ़ है। इनमें से कई पाठ और ऑडियो संसाधन की दृष्टि से कम-संसाधनीय हैं, और इनमें प्रायः प्राकृतिक भाषा अनुप्रयोगों जैसे वर्तनी सुधारक, शब्दभेद (पार्ट ऑफ़ स्पीच) टैगर आदि का अभाव रहता है। परंतु, ऑडियो रिकॉर्डिंग, नाटक, कहानियाँ और शब्दकोशों के रूप में, इन भाषाओं में बड़ी संख्या में डिजिटल संसाधन हैं (मशीन-अपठनीय)। ऐसी सामग्रियों का एक प्रमुख भण्डार है 'लैटिन अमेरिका की देशी भाषाओं का अभिलेखागार' (आईला, AILLA)।[1] आईला के संग्रह में मौजूद दस्तावेजों में से एनएलपी शोधकर्ताओं के लिए विशेष रूप से दिलचस्प भाषाई सामग्रियां हैं - व्याकरण, शब्दकोश, नृवंशविज्ञान (एथनोग्राफी) और फील्ड नोट्स, जो एनएलपी अनुप्रयोगों

[1]LLILAS बेनसन लैटिन अमेरिकन स्टडीज़ एंड कलेक्शन एंड टेक्सस विश्वविद्यालय (ऑस्टिन) द्वारा एक संयुक्त प्रयास।

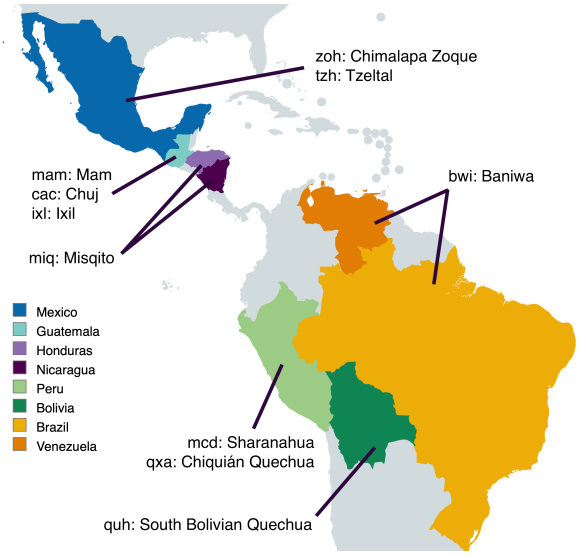

Figure 1: अपने सर्वेक्षण और उसमें निहित शोधकर्ताओं और भाषावैज्ञानिकों के लिये कार्यप्रवाह सिफ़ारिशों की भूमिका दर्शाने के लिये हम इस चित्र में मध्य और दक्षिण अमेरिका से १० ऐसी देशी महामेरिकी भाषाओं को चिह्नित करते हैं जिनमें कई अडिजिटलीकृत संसाधन मौजूद है।

और प्रकाशिक संप्रतीक अभिज्ञान (ओसीआर) के लिए प्रशिक्षण डेटा के रूप में काम कर सकते हैं। सैकड़ों डेटासेटों के ऐसे निक्षेपागार के डिजिटल संस्करण प्रकाशित करने से अमूल्य भाषाई सामग्रियों को संरक्षित किया जा सकता है और एनएलपी में अनुसंधान को गति दी जा सकती है। आधुनिक ओसीआर ऐसे दस्तावेजों से पाठ निकाल सकता है, लेकिन इसके लिए सटीक खाका संसूचन और पश्चात-प्रसंस्करण की ज़रूरत होती है ताकि निकाले गए पाठ को अनुप्रवाह (डाउनस्ट्रीम) एनएलपी कार्यों के लिए इस्तेमाल योग्य बनाया जा सके (Bustamante et al., 2020)। ओसीआर एक सुस्थापित क्षेत्र है, जिसकी प्रगति मुख्यतः कंप्यूटर दृष्टि (विज़न) में हुए नवाचारों से प्रेरित है। हाल ही में, एनएलपी संबंधित शोधों और परियोजनाओं में कम संसाधनीय भाषाओं के लिए संसाधन-निर्माण हेतु ओसीआर के उपयोग में बढ़ौतरी हुई है (Ignat et al., 2022a) । व्यापक अनुप्रयोगों के लिए ओसीआर के निर्माण और उपयोग पर कई बढ़िया सर्वेक्षण और ट्यूटोरियल उपलब्ध हैं (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020) , लेकिन कम संसाधनीय भाषाओं के ओसीआर के लिये एक विशिष्ट सर्वेक्षण की सख़्त कमी है। इसलिए, इस पेपर का

भाषा	693-3	परिवार	मुख्य देश	भाषाभाषी	पृष्ठ	अडिजिटलीकृत संसाधन
दक्षिणी बोलीवियाई केचुआ	QUH	केचुआई	बोलीविया	1.6M	216	Kalt (2016)
मिश्किितो	MIQ	मिसुमालपान	निकारागुआ, होनदूरास	150K	61	Bermúdez Mejía (2015)
माम	MAM	मायाई	ग्वातेमाला	600K	144	England (1972-1985)
चुज	CAC	मायाई	ग्वातेमाला	60K	564	Hopkins (1964)
चिमालापा ज़ोके	ZOH	मिक्से-ज़ोकियाई	मेक्सिको	<10K	3744	Johnson (2000-2005)
चिकुआई केचुआ	QXA	केचुआई	पेरू	100K	29	Proulx (1968)
शरानाहुआ	MCD	पनोआई	पेरू	<10K	209	Déléage (2002)
त्सेलताल	TZH	मायाई	मेक्सिको	600K	38	Kaufman (1960-1993)
बानीवा	BWI	माईपुरियाई	ब्राज़ील, वेनेज़ुएला	12K	310	Wright et al. (2000)
इक्सिल	IXL	मायाई	ग्वातेमाला	120K	2	Adell et al. (2016)

Table 1: महामेरिका की उन १० देशी भाषाओं को यहाँ प्रस्तुत किया गया है जिन्हें उच्च गुणवत्ता वाले ओसीआर की सख़्त आवश्यकता है। हम उनके ISO 693-3 कोड, मुख्य बोलने वाले देश, भाषाभाषियों की संख्या और उस संसाधन के कुछ संदर्भ शामिल करते हैं जिसके लिए डिजिटलीकरण की आवश्यकता है। कुल मिलाकर, इस भंडार में ५३१७ पृष्ठ शामिल हैं जिन्हें अनुलेखित किया जाना है, जिन्हें यदि डिजिटलीकृत/अनुलेखित किया जाए तो कई अनुप्रवाह एनएलपी कार्यों को प्रशिक्षित करने के लिए पर्याप्त मात्रा में डेटा निकाला जा सकता है।

उद्देश्य इस अंतर को मिटाना है और शोधकर्ताओं और भाषा समुदायों को कम संसाधनीय विन्यासों में उच्च गुणवत्तापूर्ण डिजिटलीकरण के लिए आवश्यक तकनीकों और अनुकूलनों से परिचित कराना है। संक्षेप में, यह आलेख निम्नलिखित योगदान देता है:

1. १० अमेरिकी देशी भाषाओं में अडिजिटलीकृत संसाधनों पर प्रकाश डालता है(§2).
2. कम संसाधन विन्यासों और भाषाओं के लिए ओसीआर का पहला संक्षिप्त सर्वेक्षण (§3).
3. कम संसाधनीय भाषाओं के डिजिटलीकरण के पैमानीकरण (स्केलिंग) से जुड़ी मुख्य खुली चुनौतियों पर चर्चा(§4).
4. संपूर्ण संसाधन क्यूरेशन और डिजिटलीकरण पाइपलाइन पर शोधकर्ताओं, भाषावैज्ञानिकों और भाषा समुदायों के लिए सिफारिशें (§5).

2 अडिजिटलीकृत डेटा

पिछले दशक में, कई शोधकर्ताओं, भाषावैज्ञानिकों , और कॉन्सॉर्सियमों ने देशी वक्ताओं और भाषाई समुदायों के साथ साझेदारी बनाकर कई डेटासेटों का निर्माण किया है जिनमें डिजिटलीकृत पाठ, ऑडियो, अनुलेखन (ट्रांसक्रिप्शन), अनुवाद, कथाएँ, इत्यादि शामिल हैं। इनमें से कुछ संसाधन मशीन-पठनीय न होने के बावजूद एनएलपी के नज़रिए से बेहद लाभदायक हैं जैसे बहुभाषी कोष (लेक्सिकोन), उच्चारण गाइड, विविध प्रक्षेत्रों जैसे कथाएँ, निबंध, नाटक, खबर, भाषाविज्ञान इत्यादि से केवल पाठ्यांश (प्लेन टेक्स्ट)। सभी (६०००+) कम संसाधनीय भाषाओं में इन संसाधनों की एक व्यापक सूची बहुत ही मूल्यवान होगी ज़रूर, पर इस पेपर की परिधि या विषय-क्षेत्र से बाहर है, इसीलिए हम केवल १० अमेरिकी देशी भाषाओं में मौजूद कुछ संसाधनों को दर्शाने पर ध्यान केंद्रित करते हैं। टेबल1 में दर्शाई गई चयनित भाषाओं में और सैकड़ों और अमेरिकी देशी भाषाओं में आईला संग्रह में मशीन-अपठनीय प्रतिबिंबीय संरूप में कई पाठ्य कॉर्पस मौजूद हैं। चयनित १० भाषाएँ में ही एक साथ ५००० से भी अधिक अडिजिटलीकृत पृष्ठ इस संग्रह में हैं। और अधिकांश पृष्ठों में बहुभाषी पाठ डेटा है जिसे श्रेष्ठ-गुणवत्ता के निष्कर्षण की ज़रूरत है।

कम संसाधनीय भाषाओं के ओसीआर के क्षेत्र में अधिकांश कार्य ऐतिहासिक डेटा, प्रारंभिक मुद्रित पुस्तकें (Reul et al., 2017) , ताड़-पत्र पांडुलिपियाँ (Prusty et al., 2019; Sharan et al., 2021; Alaasam et al., 2019) इत्यादि के संरक्षण और डिजिटलीकरण के इर्दगिर्द रहता हैं। पहले से मौजूद निक्षेपागार (रिपॉज़िटरी जैसे पब्मेड या आर्काइव - arXiv) भी ओसीआर प्रणालियों के प्रशिक्षण के लिए व्यापक रूप से इस्तेमाल किए जाते हैं (Zhong et al., 2019; Blecher et al., 2023) , लेकिन ध्यान रखिए कि यह तरीक़ा कम संसाधनीय विन्यासों में पैमानीय नहीं है क्योंकि इन विन्यासों में आम तौर पर ऐसे बने-बनाये डेटासेट नहीं मिलते।

महामेरिका के लिये, विस्तारित लातिनी वर्णमाला के व्यापक स्वीकृति के कारण, पिछली कुछ शताब्दियों से मिले पाठ अक्सर टाइप किए गए होते हैं, लेकिन कई संग्रहों में आंशिक या पूर्ण रूप से हस्तलिखित दस्तावेज़ और अभिटिप्पण भी होते हैं। ऐतिहासिक रूप से इस्तेमाल कि गये टाइपिंग फाँटों का गूढ़वाचन करना (डिसाइफर) मुश्किल हो सकता है या फिर ऐसा हो सकता है कि वह वर्तनी सुधारों के कारणवश प्रयोग के दायरे से ही बाहर हो गये हों (Naoum et al., 2019; Klaiman and Lehne, 2021; Jiang et al., 2019) , और हस्तलेखन का तरीक़ा व्यक्ति दर व्यक्ति व्यापक रूप से बदलता हैं जिससे निष्कर्षण और भी कठिन हो जाता है (Déjean and Meunier, 2019; Alaasam et al., 2019; Sharan et al., 2021) । समय के साथ, भाषाई समुदाय शायद कोई नयी ही वर्तनी या वर्णमाला को स्वीकृत कर सकते हैं, जिसकी वजह से शोधकर्ताओं को नये कीबोर्ड और अनुलेखन प्रणालियाँ बनाने की आवश्यकता पड़ती है ताकि डिजिटलीकृत कॉर्पस समुदाय द्वारा पठनीय हों (Ri-

jhwani et al., 2023) । इन संसाधनों के डिजिटलीकरण से भाषाई अनुसंधान, भाषा मॉडल प्रशिक्षण, अनुवाद प्रणाली, पीओएस टैगर्स आदि अधिक अभिगम्य (एक्सेसिबल) हो सकेंगे। आईला संग्रह में टाइप किए गए और हस्तलिखित दोनों प्रकार के पाठों का अच्छा ख़ासा मिश्रण है। जैसा कि टेबल 1 में साफ़ है, हाइलाइट की गई भाषाओं को अपने संबंधित संसाधनों को डिजिटल बनाने के लिए निरंतर ओसीआर प्रयासों की आवश्यकता होगी। और फिर, मशीन-पठनीय पाठ होने के बाद, अनुप्रवाह एनएलपी उपकरणों का निर्माण शुरू किया जा सकता है।

ग़ौर कीजिए कि हमारे इस संक्षिप्त सर्वेक्षण पेपर में हमारा यह उद्देश्य नहीं है कि हम इन विशिष्ट किताबों का डिजिटलीकरण करें - उसके लिए तो एक अलग और ध्यानपूर्वक रचित शोध की ज़रूरत होगी क्योंकि यह निश्चित है कि हर संसाधन के अंदर कई अनोखी चुनौतियाँ होंगी और यह अलग अलग भाषाई समुदायों से जुड़े हैं जिनकी अलग अलग इच्छाएँ और ज़रूरतें हो सकती हैं। इसीलिए यह पेपर, उन शोधकर्ताओं के लिए जो इन भाषाओं और इस प्रक्षेत्र से नावाक़िफ़ हैं, ओसीआर मॉडल बनाने की विधियों के लिए उपलब्ध विविध संसाधनों पर प्रकाश डालता है और इस स्तर के डिजिटलीकरण को साकार करने के लिए कुछ कार्यप्रवाहों की सिफ़ारिशें सामने पेश करता है।

3 ओसीआर का एक संक्षिप्त सर्वेक्षण

अब जब हम अपनी चयनित १० देशी भाषाओं में डेटा संसाधनों को देख चुके हैं (§2), हम कुछ उपयोगी और व्यावहारिक ओसीआर अनुकूलनों और नवाचारों पर ज़ोर देंगे जो कि कम संसाधनीय भाषाई डेटा के डिजिटलीकरण के लिए ज़रूरी हैं। डिजिटलीकरण पाइपलाइन की तकनीकों को हम चार मुख्य भागों में विभाजित कर कवर करेंगे: डेटा और मॉडल की तैयारी, सक्रिय प्रशिक्षण, विकोडन या जनन, और पूर्व-प्रसंस्करण। आगे आने वाली परिचर्चा की भूमिका बांधने के लिए, हम एक उदाहरण डेटासेट$\mathcal{C}$को परिभाषित करेंगे, जिसमेKपन्ने होंगे, और जहाँp_iअलग अलग पन्नों को रिप्रेजेंट* करेगा। L हर$p_i \in C$ लेबल युग्मों को निरूपित करेगा(जहाँ हरl_iपन्ने p_i के थल-सत्य (ग्राउंड ट्रुथ) शब्दों और वर्णों को निरूपित करेगा।

$$\mathcal{C} = \{p_i\}_{i=1}^{K}; \mathcal{L} = \{l_i\}_{i=1}^{K}$$

एक ओसीआर प्रयोगात्मक विन्यास के लिए, हमारे पास प्रायः चार अलग डेटासेट होंगे:C_{pretrain} (अलेबलिकृत), C_{train} (L_{train} लेबलों के साथ), C_{val} (L_{val} लेबलों के साथ प्रशिक्षण के दौरान मूल्यांकन के लिए इस्तेमाल किए गये वैधीकरण (वैलिडेशन) /विकास समुच्चय), C_{test} (L_{test}लेबलों के साथ मॉडल का निष्पादन प्रतिवेदित करने के लिए).

3.1 तैयारी: कुछ बुनियादी तकनीकें

डेटा संवर्धन कम संसाधनीय देशी भाषाओं में डेटा की कमी के कारण, डेटा संवर्धन किसी भी डिजिटलीकरण पाइपलाइन का पहला कदम होना चाहिए, ताकि छोटे लेबलिकृत गोल्ड डेटा की उपयोगिता बढ़ाई जा सके (Shorten and Khoshgoftaar, 2019)। किसी ओसीआर प्रणाली के लिए, इसका यह मतलब होगा कि प्रतिबिंबों को ख़ुद कई रूपांतरणों से गुज़ारना होगा जैसे कि विषमन (स्केविंग), द्विआधारीकरण (बाइनराइज़ेशन), पैमानीकरण, दृश्यांकन (क्रॉपिंग), अस्पष्टीकरण (ब्लोरिंग), इत्यादि, यह सुनिश्चित करने के लिए कि आख़िरी/फाइनल मॉडल इस तरह की सभी विविधताओं को इन-द-वाइल्ड पकड़ सके और उनसे पाठ निष्कर्षित करने में सक्षम रहे। डेटा संवर्धन साहित्य में सुअध्ययनित विषय है(Liu et al., 2018; Khan et al., 2021) और इसे ओसीआर पाईपलाइनों में समावेशित करने से इनकी दृढ़ता (रॉबस्टनेस) और निष्पादन में बढ़ोतरी देखी गई है जो कि छोटे प्रशिक्षण सेट के श्रेष्ठ इस्तेमाल से होता है(Storchan and Beauschene, 2019; Namysl and Konya, 2019)।

और बारीकी से देखें तो, संवर्धन प्रचालनों का एक समुच्चय, $O = \{o_1, o_2, .., o_j\}$जहाँ j प्रचलनों की संख्या को दर्शाता है जो कि हर प्रतिबिंब p_iपर लागू किए जा सकते हैं।oद्विआधारीकरण, ग्रेस्केल, गॉसियन अस्पष्टता/ब्लर, दृश्यांकन इत्यादि जैसे फलनों को निरूपित करता है। C_{train} को समुच्चयOके अंदर मौजूद प्रचालनों के किसी भी संयोजन से संवर्धित किया जा सकता है, जिससे एक नया समुच्चय $C_{\text{train-aug}}$बन कर सामने आयेगा, जो कि हमारा नया विस्तारित प्रशिक्षण कार्पस बनेगा। हर नये विस्तारित पन्ने $p_{i,j} = o_j(p_i)$ के लिए $l_i \in L_{\text{train}}$उसका लेबल होगा।

सामान्य अनलेबलिकृत डेटा के साथ पूर्व-प्रशिक्षण ऐसे डेटा के लिए जो कि लेबेलिकृत नहीं है, स्व-पर्यवेक्षण पूर्वप्रशिक्षण तकनीकें आम तौर पर इस्तेमाल की जाती हैं ताकि नेटवर्क को बेहतर ढंग से प्रारम्भीकृत किया जा सके(Li et al., 2023; Bugliarello et al., 2021)। कोडक-विकोडक माडलों के मामले में, पूर्वप्रशिक्षण को दोनों घटकों पर पृथकतः लागू करने को सफलतापूर्वक पाया गया है (Lyu et al., 2022; D'hondt et al., 2017),ख़ास तौर पर जब उपलब्ध अलेबलिकृत प्रतिबिंबों और पाठों की संख्या बहुत ज़्यादा हो। इसी तरह, आंतरिक पूर्वप्रशिक्षण समुच्चयC_{pretrain} के मामले में थल-सत्य पाठ लेबल उपलब्ध नहीं होते हैं। तो, इस पूर्वप्रशिक्षण समुच्चय की प्रतिबिंबों को ओसीआर मॉडल के पूर्वप्रशिक्षण के लिए इस्तेमाल किया जा सकता है, और प्रथम-पारण पाठ को पश्चात-सुधार मॉडल के पूर्वप्रशिक्षण में समावेशित किया जा सकता है (कुछ अधिगमित अपशोरीय नियमों के साथ) (Rijhwani et al., 2020)।

संबंधित भाषाओं से स्थानांतरण अधिगम कुछ चुनिंदा अल्पसंख्यक एवं कम-संसाधनीय भाषाओं के लिये, यह पाया गया है कि एक मूल प्रणाली जिसे संबंधित भाषा या फिर संबंधित वर्णमाला पहचानने के लिये प्रशिक्षित किया गया है, वह आम तौर पर अनुप्रवाह में निष्पादन में बढ़ौतरी का कारण बनती है।(Lin et al., 2019; Zhuang et al., 2021; Rijhwani et al., 2019)। उदाहरण के रूप में, हमारी चयनित १० अमेरिकी देशी भाषाओं में, केंद्रीय एवं दक्षिण अमेरिका की एक अधिक-संसाधनीय भाषा जैसे स्पेनी या पुर्तगाली, कार्पस चुनना हमारे स्थानांतरण प्रयोगों में फ़ायदेमंद साबित हो सकती है। ओसीआर प्रक्षेत्र में,

स्थानांतरण अधिगम को विकोडन चरण में कम-संसाधनीय ओसीआर निर्गम (आउटपुट) की गुणवत्ता को बढ़ाने के लिए इस्तेमाल किया गया है(Todorov and Colavizza, 2020; Jaramillo et al., 2018)। परंतु, Tjuatja et al. (2021)के शोध में स्थानांतरण अधिगम की देशी और संकटापन्न भाषाओं के लिये जाँच पड़ताल की गई और मिश्रित नतीजे पाये गये। वह अपने पेपर में कहते हैं कि अनुप्रवाह सुधार के लिए, स्थानांतरण अधिगम इतना सीधासादा भी नहीं है और इसके लिए हमें बड़ी संख्या के अलग अलग प्रक्षेत्रों से डेटा प्राप्त करने की आवश्यकता होगी।Tjuatja et al. (2021)ने भारतीय भाषाओं के लिये पाठ-संसूचन स्तर पर स्थानांतरण पर शोध किया, और सकारात्मक नतीजे पाये जब ऐसी भारतीय भाषाओं से स्थानांतरण किया जाये जो देखने में यानी अपनी वर्णमाला में एक जैसी दिखती हैं, चाहे वह अलग अलग भाषा परिवारों से ही क्यों न हो।

3.2 प्रशिक्षण: जल्दी और बेहतर सीखें

ओसीआर सिस्टम प्रशिक्षित करने के लिए, कम संसाधनीय विन्यासों में पर्यवेक्षित तकनीकों को प्राय: पसंद* किया जाता है। अपर्यवेक्षित तकनीकों ने भी हाल ही में कुछ सकारात्मक निशान दिखाए हैं(Gupta et al., 2021; Dong and Smith, 2018; Garrette and Alpert-Abrams, 2016), पर उनके प्रशिक्षण के लिए कई गुना ज़्यादा बड़े डेटासेटों की ज़रूरत पड़ती है। चूँकि हमारा ध्यानकेंद्र कम-संसाधनीय देशी भाषाओं हैं, हम अपनी परिचर्चा को पर्यवेक्षित तकनीकों तक सीमित रखते हैं। इस विन्यास में, आम तौर पर दो विकल्प होते हैं - ऑफ-द-शेल्फ* सिस्टम जैसे गूगल विज़न, टेसरेक्ट इत्यादि, या फिर स्क्रैच* से प्रशिक्षण करना। अमेरिकी देशी भाषाओं के लिये, ऑफ-द-शेल्फ* सिस्टम एक बहुत ही अच्छा स्टार्टिंग पॉइंट* दे सकते हैं(Rijhwani et al., 2023),और चूँकि यह इस पेपर का ध्यानकेंद्र हैं, इसीलिए हम ऐसे सिस्टमों से प्राप्त प्रथम-पारण ओसीआर के ऊपर की गयीं प्रशिक्षण तकनीकों पर चर्चा करेंगे। पश्चात-सुधार ओसीआर सिस्टमों द्वारा पाठ निष्कर्षण में की गई ग़लतियों को सुधारने का उद्देश्य रखता है, और कम-संसाधनीय भाषाओं के लिये बहुत ही मूल्यवान है। पश्चात-प्रसंस्करण क़ीमती इसीलिए भी है क्योंकि वह प्रथम-पारण ओसीआर सिस्टम के बारे में कम से कम असम्पशन* बनाता है (जो की बहुत काम का होता है जब सिस्टम कमर्शियल* या क्लोज्ड-सोर्स* हो) और उसके बरक्स आउटपुट* की गुणवत्ता बढ़ाने पर ध्यान देता है (Kolak and Resnik, 2005) ।

प्रथम-पारण ओसीआर प्रथम-पारण के लिए, एक श्रेष्ठ-गुणवत्तीय ओसीआर सिस्टम, जैसे गूगल विज़न या टेज़रेक्ट, का इस्तेमाल प्राय: किया जाता है क्योंकि यह संकटापन्न-भाषीय दस्तावेज़ों पर अच्छे प्रदर्शन के लिए इस्तेमाल किए गए हैं(Fujii et al., 2017; Rijhwani et al., 2020)। पन्नेp_iपर ओसीआर करने से हमें प्रथम-पारण आउटपुट* f_i मिलता है जिसमें n_iबाउंडिंग बॉक्स* और हर डब्बे में पाठ a। हरxके अंदर बाउंडिंग बॉक्स* के कूर्डिनेट्स* का समुच्चय होता है, और उसकी करेस्पोंडिंग* स्ट्रिंग* उस डब्बे के भीतर के पाठ को रिप्रेज़ण्ट* करती है।

$$f_i = [(x_1, a_1), (x_2, a_2), .., (x_{n_i}, a_{n_i})]$$

पाठ सुधार एक आदर्श पश्चात-ओसीआर पाठ सुधार कलनविधि ओसीआर विधि के आउटपुट पाठ के त्रुटि बंटन को मॉडल कर उसे व्यवस्थितत: सुधारेगी (Berg-Kirkpatrick et al., 2013; Schulz and Kuhn, 2017)। देशी भाषाओं के दस्तावेज़ों को डिजिटलीकृत करते समय यह एक बहुत ही मूल्यवान उपकरण साबित होता है क्योंकि ओसीआर पाइपलाइन का विकोडन भाषा मॉडल संकटापन्न और देशी भाषाओं की कम-संसाधनीय प्रकृति के कारण आम तौर पर बहुत ही कम गुणवत्ता का होता है। उन सभी डिजिटलीकरण प्रयासों को जिन्हें हमने चिह्नित किया है और अन्य प्रयासों में भी, यह आम बात है कि पाठ-आधारित सेमी-स्वसंचालित* या मानव पश्चात-सुधार किया जाए (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021) । हर प्रथम-पारण पन्ने f_iके लिए, हम एक सुधारित* पन्ने को आउटपुट करेंगे:

$$q_i = [(x_1, b_1), (x_2, b_2), .., (x_{n_i}, b_{n_i})]$$

जहाँ x का मतलब है प्रथम-पारण के डब्बे, और b का मतलब है उसका कॉरेस्पिंडिंग सुधारित पाठ। मानव पश्चात-सुधार में, एक अभिटिप्पणकर्ता (अधिमानत: भाषा को बोलने-समझने वाला), प्रथम-पारण ओसीआर आउटपुट को संशोधित करेगा ताकि वह करेस्पोंडिंग थल सत्य पाठ, जैसा कि प्रतिबिंब में दिखता है, से मेल खा सके। सेमी-स्वसंचालित* विन्यासों में, कई सुसंगत ओसीआर त्रुटियों को एक छोटी सी संख्या के सुधारों से ही पहचाना जा सकता है और फिर उन्हें स्वसंचालित: बचेकुचे प्रथम-पारण प्रागुक्तियों पर लागू किया जा सकता है ताकि अभिटिप्पणकर्ता के ऊपर काम का बोझ हल्का हो।

व्याप्ति क्रियाविधि चूँकि ओसीआर को एक जनन कार्य की तरह देखा जाता है, मॉडल के ध्यान बंटन के लिए यह ज़रूरी हो सकता है कि वह निवेश स्ट्रिंग की के अलग अलग भागों पर ध्यान दे। यह सुनिश्चित करने के लिए कि ऐसा हो, एक व्याप्ति (कवरेज) क्रियाविधि को आम तौर पर समावेशित किया जाता है(Tu et al., 2016; Mi et al., 2016)। इस क्रियाविधि ने आनुभविक रूप से ओसीआर और सीक-टू-सीक निष्पादन को काफ़ी बढ़ाया है (See et al., 2017; Rijhwani et al., 2021; Klaiman and Lehne, 2021) । कालचरणt का व्याप्ति सदिश (वेक्टर) होगा:

$$c_t = \sum_{t'=0}^{t'=t-1} \alpha_{t'}^a$$

जहाँ α_t^a कालचरण t के निवेश a के ध्यान बंटन को निरूपित करता है। यह व्याप्ति सदिश c_t भारित (वेटिड) कर अगले α_{t+1} की ध्यान अभिकलन (कंप्यूटेशन) में अंतर्वेशित किया जा सकता है, और मूल क्रॉस-एंट्रॉपी व्यय (लॉस) में ऐसे जोड़ा जा सकता है:

$$\sum_t \sum_{i=0}^{\text{len}(a)} \min(\alpha_{t,i}^a, c_{t,i})$$

विकर्ण ध्यान चूँकि प्रथम-पारण ओसीआर से पश्चात-सुधार मूलतः एक प्रतिलिपीकरण चरण है और पुनरादेशीकरण विरल होता है (Schnober et al., 2016), मॉडल मुख्यतः विकर्ण के आसपास के अवयवों के जनन पर ध्यान केंद्रित कर सकता है। इसीलिए, इस रूपावली के अंतर्गत, एक विशिष्ट त्रिज्या के बाहर की ऑफ-विकर्ण प्रविष्टियों को प्रशिक्षण व्यय में शामिल कर उन्हें और ज़्यादा भारी तरह से दंडित किया जा सकता है(Cohn et al., 2016)। इससे विकोडन चरण को सरलीकृत कर देता है और विकर्ण ध्यान परिसर में मौजूद मदों पर ध्यान अधिकतम करने के लिए मॉडल को प्रोत्साहित करता है। तो आपरिवर्तित (मॉडीफ़ाइड) व्यय फलन, कालचरण t पर एक विकर्ण परिसर d और ध्यान बंटन α के लिए होगा:

$$\sum_{t'=1}^{t-d} \alpha^a_{t,t'} + \sum_{t'=t+d}^{\text{len}(a)} \alpha^a_{t,t'}$$

यह देखा गया है कि विकर्ण ध्यान आनुभाविकतः कम-संसाधनीय भाषाओं के ओसीआर के निष्पादन में सुधार लाता है (Rijhwani et al., 2021, 2020) और इसे आसानी से ओसीआर पश्चात-सुधार मॉडल बनाने में समावेशित किया जा सकता है।

सक्रिय अधिगम डेटा लेबलिंग कम-संसाधनीय भाषाओं के लिये एक महँगा कार्य है, ख़ास तौर पर ओसीआर सुधार या प्रतिबिंब लेबलिंग जैसे अतुच्छ अभिटिप्पण कार्यों के लिए। केवल उन पन्नों को अभिटिप्पण के लिए चुनने के लिए जो ओसीआर मॉडल की सबसे ज़्यादा सहायता करेंगे, एक व्यवस्थित रूपावली 'सक्रिय अधिगम' इस्तेमाल किया गया है (Settles, 2012) । कम-संसाधनीय ओसीआर प्रक्षेत्र और खाका विश्लेषण के लिए, सक्रिय अधिगम आनुभाविकतः बहुत मूल्यवान साबित हुया है (Reul et al., 2018; Shen et al., 2022; Monteleoni and Kaariainen, 2007; Abdulkader and Casey, 2009; Gupta et al., 2016)। यह हमें यह चुनवाने में मदद कर सकता है कि C_{train} में C_{pretrain} के कौनसे हिस्से को डाला जाए, जो कि यह क्वेरी-बाय-कमेटी तकनीक से करता है। क्वेरी-बाय-कमेटी कई विद्यार्थी मॉडल C_{train} पर प्रशिक्षित करती है और फिर हर मॉडल C_{pretrain} से लिये गये एक अलेबलीकृत नमूनों के समुच्चय पर अपना मत/प्रागुक्ति देता है। नीचे दी गयी समीकरणों में, $\text{uq}(\cdot)$ प्रागुक्तियों की एक सूची में अद्वितीय वर्णों की संख्या गिनता है, M स्वतंत्र रूप से प्रशिक्षित माडलों को निरूप्रित करता है (कुल मिलाकर m),s_v C_{pretrain} के v^{th} वाक्य को निरूप्त करता है, और $V = \text{len}(C_{\text{pretrain}})$ ।

$$\text{ag}_{s_v} = \text{uq}([M_1(s_v), M_2(s_v), ..., M_m(s_v)])$$

$$v* = \text{argmax}^V_{v=0}(\text{ag}_{s_v})$$

प्रतिदर्श $v* \in C_{\text{pretrain}}$ वो प्रतिदर्श है जिसपर मॉडल सबसे ज़्यादा असहमत हैं और इसीलिए इसे प्रशिक्षण समुच्चय C_{train} में सक्रिय रूप से डाला जाता है (उच्चतम असहमति का सिद्धांत) क्योंकि इसे मानव अभिटिप्पण की सबसे ज़्यादा आवश्यकता है और इससे ओसीआर मॉडल सबसे ज़्यादा बेहतर हो पाएगा(Settles, 2012)।

3.3 विकोडन: जनन करें कि नहीं?

इस उपभाग में, हम कम-संसाधनीय विन्यासों में ओसीआर विकोडन को बेहतर बनाने की कुछ हाल ही में प्रस्तावित और आनुभाविकतः कामयाब रणनीतियों पर परिचर्चा करेंगे।

प्रतिलिपि प्रणाली चूँकि विकोडन चरण पर यह संभावित है कि सुधारित पाठ में ज़्यादातर पाठ निवेश (इनपुट) के समरूप होगा, यह उपयोगी दिखाया गया है (Gu et al., 2016) कि हम विकोडन के लिए दो प्रायिकता बंटन रखें - प्रति और जनन। विकोडन के समय, मॉडल यह चुन सकता है, कि ध्यान बंटन से प्रतिचयन किया जाये (P) या फिर निर्गम/आउटपुट को जनन द्वारा जना जाये (See et al., 2017; Sutskever et al., 2014)।

$$P(y_t) = \sum_{t'=0}^{t} \alpha_{t,t'}$$

इससे ओसीआर के वर्ण और शब्द त्रुटि दरें २-५ गुना तक कम-संसाधनीय विन्यासों में कम हो सकते हैं (Rijhwani et al., 2020; Gu et al., 2016) । Krishna et al. (2018) भी संस्कृत ओसीआर के लिये एक प्रतिलिपि प्रणाली का इस्तेमाल करते हैं और इसके ज़रिये मूल मॉडल की अपेक्षा में उन्हें क़रीबन १०% पॉइंट का लाभ हुयामाँ, जिससे यह प्रमाणित होता है कि प्रतिलिपि प्रणाली को कम-संसाधनीय और देशी भाषाओं की ओसीआर पाइपलाइन में समावेशित करना बहुत फ़ायदेमंद हो सकता है। प्रति-प्रायिकता को हर कालचरण पर एक $p \in (0, 1)$ के आधार पर भारित किया जा सकता है, और यह संदर्भ सदिश, विकोडक स्टेट, और पिछले कालचरण में विकोडक प्रायिकता के एक भारित जोड़ से जनित किया जा सकता है। तो हमें निम्नलिखित प्रतिजनन प्रायिकता, विशिष्ट कालचरण t और निर्गम स्ट्रिंग y के लिए प्राप्त होती है:

$$p(y_t) = (1 - p_{\text{copy}}) * P(y_t) + p_{\text{copy}} * P_{\text{copy}}(y_t)$$

कोशीय विकोडन पिछले प्रशिक्षण चरण के स्व-प्रशिक्षण के उत्पन्न होने वाले शोर (यानी प्रथम-पारण की त्रुटियों का सुदृढ़ होना) का मुक़ाबला करने के लिए, कोशिया अनुकूलनों को सफलतापूर्वक विकोडन चरण में परिचित किया गया ताकि प्रागुक्ति की गुणवत्ता को बढ़ाया जा सके (Schulz and Kuhn, 2017; Rijhwani et al., 2021) । इस प्रस्तावित उपाय ने यह दर्शाया है कि यह विकोडन को आनुभविक फ़ायदा पहुँचाता है क्योंकि यह यह कल्पना करता है कि शब्द के सही रूप ज़्यादा होंगे (यह मान ले कि ओसीआर त्रुटियाँ असंगत हैं) और निर्गम को ऐसे प्रेक्षित रूपों की ओर अभिनति कर देता है।

3.4 मूल्यांकन: सुधार को कैसे मापें?

पूर्वानुमान अंकन और मूल्यांकन मानक जब हम एक ओसीआर प्रणाली को बिलकुल शून्य से बनाते हैं, तो माध्य-औसत-परिशुद्धता (मीन-एवरेज-प्रिसिशन, mAP) और प्रतिच्छेदन-पर-संघ (इंटरसेक्शन ओवर यूनियन, IoU) दो सबसे ज़्यादा इस्तेमाल किए जाने वाले मापदंड हैं जिनसे सीमक बक्सों की गुणवत्ता मूल्यांकित की जा सकती है। प्रागुक्तित सीमक बक्सों $P = \{x_1, x_2, ..., x_e\}$ के लिए, शो-

धकर्ता आम तौर पर IoU का इस्तेमाल सभी बक्सों के युग्मों पर करते हैं ताकि सीमक बॉक्स प्रागुक्ति और संदर्भ युग्मों की एक श्रेणीबद्ध सूची जनित करी जा सके (Girshick, 2015; Prasad et al., 2019; Prieto and Vidal, 2021) । फिर, IoU देहलियों के एक परिसर से एक कन्फ़्यूशन मैट्रिक्स जनी जा सकती है, जिससे हमें परिशुद्धता और प्रत्याह्वान के युग्म उस देहली के लिए मिल सकते हैं। हर देहली के लिए इन युग्मों का आलेखन कर, हमें एक परिशुद्धता-प्रत्याह्वान वक्र मिलता है, जिसके नीचे का क्षेत्रफल औसत-परिशुद्धता (AP) कहलाता है। हम हर संदर्भ बक्से x_e के लिए AP पा सकते हैं, और इन सभी की औसत निकालने से हमें उस पन्ने की mAP मिल जाएगी। यह प्रागुक्तियों P की वास्तविक संदर्भ लेबलों के साथ संरेखण की गुणवत्ता की ओर इशारा करता है।

परंतु, कई महामेरिकी देशी भाषाओं के लिये, ऑफ-द-शेल्फ और व्यावसायिक प्रणालियाँ एक ठीक-ठाक प्रथम-पारण प्रागुक्ति पैदा करेंगे क्योंकि वह लातिनी वर्णमाला के ही विस्तारित रूपों का इस्तेमाल करती हैं (Rijhwani et al., 2020) । इस केस में, मूल्यांकन को दो पाठ स्ट्रिंगों को मिलाना होगा: प्रागुक्ति और गोल्ड-संदर्भ। इसके लिए, वर्ण त्रुटि दर (करैक्टर एरर रेट, CER) और शब्द त्रुटि दर (वर्ड एरर रेट, WER) दो सबसे ज़्यादा लोकप्रिय मूल्यांकन मापदंड हैं। कुछ भाषाओं के लिये, दोनों CER और WER हो सकता है कि निर्देशात्मक न हों - जैसे बहुसंश्लेषणात्मक भाषाओं के लिये जहाँ शब्दावली का एक बड़ा हिस्सा मूल्यांकन के समय अनदेखा होगा, वर्ण-स्तर पर त्रुटि दर ओसीआर निष्पादन को आंकने के लिये बेहतर मापदंड दिखाई गई है (Rijhwani et al., 2023)।

$$\text{CER} = \frac{s_c + d_c + i_c}{n_c}; \text{WER} = \frac{s_w + d_w + i_w}{n_w}$$

जहाँ s, d, और i, वर्ण c और शब्द w के स्तर पर संदर्भ पाठ की अपेक्षा में (जिसमें n शब्द/वर्ण हैं), प्रतिस्थापन, विलोपन और निवेशन निरूपित करते हैं।

व्ययफलन (लॉस फलन) अगर आप एक ऑफ-द-शेल्फ़ प्रणाली का इस्तेमाल कर रहे हैं प्रथम-पारण निष्पादन पाने के लिए, तो आप जैसे शोधकर्ताओं को केवल एक पश्चात-सुधार मॉडल प्रशिक्षित करने की ज़रूरत पड़ेगी। इस केस में, एक क्रॉस-एंट्रॉपी व्यय अतिआवश्यक है, कई अन्य अनुकूली व्ययफलनों के अतिरिक्त जैसे विकर्ण और व्याप्ति व्ययफलन जिनके बारे में§3.2 परिचर्चा की गई है (Cohn et al., 2016; Tu et al., 2016) । इन व्ययों के संयोजन को इष्टतमित करने के लिए, कई प्रचलित इष्टतमक (ऑप्टिमाइज़र) जैसे स्टॉकस्टिक ग्रेडिएंट डिसेंट (एसजीडी) या ऐडम काफ़ी इस्तेमाल किया जाते हैं (Rijhwani et al., 2020) । ऐसी परिस्थितियों में जहाँ ओसीआर प्रणाली को मूल से प्रशिक्षित करना पड़े, वहाँ पर प्रति-पिक्सेल सिग्माइड या सॉफ़्टमैक्स व्ययों को कार्यरत किया जाता है क्योंकि साधारण मॉडल जैसे मास्क-आरसीएनएन (Mask R-CNN) या फ़ास्ट-आरसीएनएन (Fast R-CNN) पिक्सेल-स्तर की प्रागुक्तियाँ देती हैं (Girshick, 2015; He et al., 2017)। अगर नेटवर्क की अलग अलग शाखाएँ अभिज्ञान कार्य के

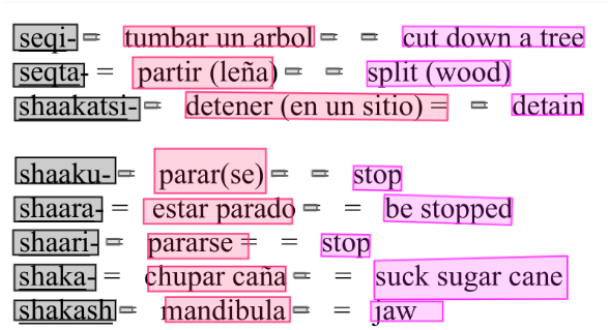

Figure 2: आईला संग्रह (§2) से लिया गया एक चिकुईआन केचुआ (स्पेनी और अंग्रेज़ी में बहुभाषी) पश्चात-सुधारित ओसीआर दस्तावेज़ । यहाँ, अभिटिप्पणकर्ता ने संसूचित सीमक बक्सों को पुनः समायोजित किया है, नये बक्सों में पाठ्य त्रुटियों को सुधारा, और तीनों भाषाओं के बक्सों को भाषानुसार अलग अलग रंगा।

अलग अलग पहलुओं का विश्लेषण करती हैं और उनपर प्रागुक्तियाँ देती हैं, तो अलग अलग व्ययफलनों को अभिकलनित किया जाता है, और कुल व्यय ऐसे कस में इन सभी के एक उत्तल (कॉन्वेक्स यानी जिसका तल उठा हो) संयोजन से अभिकलनित किया जाता है (Prusty et al., 2019)।

4 खुली चुनौतियाँ

खाका/अभिन्यास संरक्षण ओसीआर साहित्य में अनसुलझे मुद्दों में से एक अतिप्रमुख समस्या है खाका (स्ट्रक्चर) संरक्षण। ओसीआर उपकरण, ख़ास कर के वह जो ऑफ-द-शेल्फ़ हैं, किसी पन्ने के खाके को सटीक ढंग से संरक्षित करने में शायद सक्षम न हों, और कुछ हद तक पश्चात-सुधार संरेखण की दरकार हो सकती है (Tafti et al., 2016; Rijhwani et al., 2020) । हो सकता है कि संसूचित सीमक बक्से, जैसा कि इंसानी निरीक्षण से अपेक्षित हो वैसे तर्कसंगत खाके में न बटे हों। इसका यह मतलब होगा कि शोधकर्ताओं को ओसीआर निष्पादन पाने के बाद कुछ हद तक संरेखण करना पड़ सकता है (Xie and Anastasopoulos, 2023) , या फिर ओसीआर मॉडल लागू करने से पहले (Ignat et al., 2022a), या फिर हर प्रतिबिंब को अलग अलग पंक्ति-स्तरीय प्रतिबिंबों में क्रॉप कर के (वाणिज्यिक प्रणालियों का इस्तेमाल यदि कोई शोधकर्ता कर रहा है तो यह विकल्प आर्थिक रूप से अव्यावहारिक हो सकता है)। देशी भाषाओं के संसाधन-सृजन के दृष्टिकोण से, अंतिम निष्पादन में खाका संरक्षण बेहद ज़रूरी है, इसीलिए हम शोधकर्ताओं को यह सलाह देते हैं कि वह इस मुद्दे को संबोधित करने के लिए शुरुआत से ही अपने प्रयोगों के डिज़ाइन पर ध्यान से विचार करें।

हमारी जानकारी के अनुसार, जबकि पिछले शोधकर्यों ने पहले चरण के रूप में खाके का पता लगाने पर ध्यान केंद्रित किया है (Bustamante et al., 2020) , इसे पूर्वप्रसंस्करण चरण के रूप में नहीं जाँच गया है, मुख्य रूप से थल-सत्य संरचनात्मक डेटा की कमी के कारण। इससे पहले, दो बड़े अध्ययनों (Blecher et al., 2023; Zhong et al., 2019) ने मौजूदा बड़े-पैमानाई कॉर्पोरा जैसे आर्काइव का इस्तेमाल किया है बड़े-पैमानाई थल सत्य (मूल या सोर्स कोड) को

निष्कर्षित करने के लिए; लेकिन, कम-संसाधनीय भाषाओं के संसाधन-सृजन प्रयासों के लिए यह पैमानीकरण योग्य नहीं है। एसे खाका पश्चात-सुधार मॉडल का निर्माण करने के लिए, अभिटिप्पणकर्ताओं को न सिर्फ़ ओसीआर पाठ को सुधारना होगा बल्कि प्रथम-पारण ओसीआर आउटपुट को किसी तरह के ग्राफिकल यूजर इंटरफ़ेस (जैसा कि फिगर 2 में दर्शाया गया है) में उसे संरचनात्मक रूप से भी सुधारना होगा। इसके लिए पैमानीकरण, अनुवाद, मर्गिंग, या फिर सिमक बक्सों को तोड़ने के कार्य करने पड़ सकते हैं, यह सुनिश्चित करते हुए कि हर बक्से के नये कोआर्डिनेट के अनुसार ही उसके भीतर का पाठ भी संगत में हो। ऐसे कार्य का ढाँचा इस तरह तैयार किया जा सकता है: हर पाठ-सुधारित पन्ने q_i के लिए, हम एक सुधारित पन्ना आउटपुट करेंगे

$$r_i = [(y_1, c_1), (y_2, c_2), .., (y_{m_i}, c_{m_i})]$$

जहाँ m_i पश्चात-सुधार के बाद नये सीमक बक्सों की संख्या निरूपित करेगा (हो सकता है कि यह n_i से भिन्न हो)। हम मानव-सुधारित r_i को थल सत्य पाठ और खाका मानेंगे। ध्यान दीजिये कि चाहे यह चरण मुख्यतः खाके को परिवर्तित करता है, इसके अंतर्गत उन प्रथम-पारण सीमक बक्सों के प्रारंभिक सुधारित पाठ को भी स्थानांतरित होगा (b_i, b_{i+1} , इत्यादि) जो कि सुधारित बक्से y_i के अधीन थे, और इसीलिए, पाठों को अब c_i लेबल दिया गया है। परंतु, चूँकि संरचनात्मक पश्चात-सुधार थल सत्य डेटा पाना बहुत खर्चीला हो सकता है, शोधकर्ता थल सत्य को किसी खाका सुधार मॉडल से भी स्वचालित रूप से पा सकते हैं, और इस स्वचालित रूप से निकाले गए वांछित लेआउट के अनुकूल होने के लिए फिर श्रेष्ठतम प्रथम-पारण ओसीआर प्रणाली से निकले आउटपुट को पश्चात-सुधार कर सकते हैं।

असामान्य वर्ण, फ़ॉन्ट और शब्द वर्तमान वर्तनी परंपराओं पर प्रशिक्षित आधुनिक एलएम के साथ ऐतिहासिक वर्तनी भिन्नताओं की मॉडलिंग करना, विकोडन चरण के दौरान चुनौतीपूर्ण साबित हो सकता है (Poncelas et al., 2020)। मुद्रणालय युग के ऐतिहासिक दस्तावेजों से बेहतर पाठ निष्कर्षण पर किए गए कार्य के परिणामस्वरूप लोकप्रिय रूप से प्रयुक्त अपर्यवेक्षित (अनसुपर्विज़्ड) ऑक्यूलर मॉडल का विकास हुआ (Berg-Kirkpatrick et al., 2013)। असामान्य वर्णों और टाइपफेस के प्रभाव को संतुलित करने के लिए संश्लेषणात्मक डेटा का पहले भी सफलतापूर्वक इस्तेमाल किया जा चुका है (Borenstein et al., 2023; Drobac et al., 2017) , और ऐतिहासिक दस्तावेज़ पहचान और ओसीआर के संदर्भ में दस्तावेज़ की फ़ॉन्ट शैली को स्वचालित रूप से सीखने के लिए अपर्यवेक्षित तकनीकों का इस्तेमाल किया गया है(Berg-Kirkpatrick and Klein, 2014)। हालांकि, कम-संसाधनीय प्रक्षेत्र में अनुसंधान अभी भी सीमित है और शोधकर्ताओं को यह सुनिश्चित करने की ज़रूरत होगी कि उनके फ़ॉन्ट और वर्ण समुच्चय उनके चुने हुए ओसीआर मॉडल द्वारा समर्थित हैं (यदि किसी को मूल से प्रशिक्षित किया जा रहा है) या दृश्य प्रतिनिधित्व (विसुअल रिप्रजेंटेशन) सीखने में हाल के कार्य का इस्तेमाल करके पुनर्निर्मित किए गए हैं(Srivatsan et al., 2021; Vogler et al., 2022)। ऑफ-द-शेल्फ प्रणालियों (एपीआई के माध्यम से इंटरफेस किए गए) के लिए, विशिष्ट वर्णों के लिए सीधे समर्थन शामिल करना संभव नहीं है और शोधकर्ताओं को एक पश्चात-सुधार चरण जोड़ने की आवश्यकता होगी, जिसमें आसानी से हल की जा सकने वाली त्रुटियों के लिए पश्चात-प्रसंस्करण स्क्रिप्ट का मिश्रण और प्रथम-पारण आउटपुट को सही करने के लिए समर्पित प्रशिक्षित पर्यवेक्षित मॉडल शामिल होंगे।

भाषाई विविधता एक ही समय में कई कम-संसाधनीय या देशी भाषाओं के साथ काम करने वाले शोधकर्ताओं के लिए, ऐसे एक मॉडल को प्रशिक्षित करना वांछनीय हो सकता है जो विभिन्न लेखन प्रणालियों, विशेषकों, भिन्न प्रतिबिंब गुणवत्ता और अनूठे दस्तावेज़ फ़ॉर्मेटिंग को संभालने में सक्षम हो (Joshi et al., 2020) । हालांकि एक दृष्टिकोण ऐसा भी होता है जो 'भाषा-अज्ञेय' विधियों को विकसित करता है, लेकिन पिछले काम ने दिखाया है कि व्यावहारिक रूप से ऐसे मॉडल भाषा-अज्ञेयता से बहुत दूर रह जाते हैं (Joshi et al., 2020; Bender, 2011)और केवल मुट्ठी भर भाषाओं के लिए उच्च प्रदर्शन करते हैं। उच्च ओसीआर सटीकता आमतौर पर विचाराधीन सभी कम-संसाधनीय भाषाओं के लिए वांछनीय है, और एसी स्थिति में, अलग-अलग ओसीआर या पश्चात-सुधार मॉडल को प्रशिक्षित करना सबसे अच्छा हो सकता है।

5 कार्यप्रवाह सिफ़ारिशें

इस अनुभाग में, हम सर्वेक्षित पेपरों में मौजूद सबसे कामयाब तरकीबों पर आधारित कुछ सिफ़ारिशें साझा करेंगे। कम-संसाधनीय प्रक्षेत्र में आये नये अभिकलनात्मक शोधकर्ताओं, भाषावैज्ञानिकों, और विद्यार्थियों के लिए यह एक आरंभस्थल के रूप में काम आ सकता है। हम यह मानते हैं कि यह सिफ़ारिशें, चाहे हमारे सर्वेक्षण पर आधारित ही क्यों न हों, फिर भी व्यक्तिनिष्ठ हैं और शोधकर्ताओं को अपने विशिष्ट मामलों के अनुरूप कुछ तत्वों को संशोधित करने की ज़रूरत पड़ सकती है।

भाषा और दस्तावेज़ चयन अपने सर्वेक्षण की नींव बांधने के लिए, हमने १० ऐसी भाषाओं का चयन किया, जिनके पास इजाज़ती लाइसेंस हैं, जो लैटिन वर्णमाला का उपयोग करती हैं, जिनके विशेष विश्लेषक अंग्रेजी कीबोर्ड पर उपलब्ध हैं, और जिनके दस्तावेज़ सामान्य फ़ॉन्ट में टाइप किए गए हैं। इसी तरह, अपनी रुचि की भाषाओं के लिए दस्तावेज़ चुनते वक़्त, शोधकर्ताओं को लाइसेंसिंग, विशेष कीबोर्ड की ज़रूरत, दस्तावेज़ की गुणवत्ता और लेआउट/खाका विविधता पर विचार करना चाहिए।

मूल्यांकन तकनीकें ओसीआर की गुणवत्ता मूल्यांकित करने के लिए, हम अंतिम निष्पादन यानी पाठ पर साधारण नज़र डालने की पहली सिफ़ारिश करते हैं। बशर्ते कोई संदर्भ या गोल्ड पाठ हो, इन प्रागुक्तियों की तुलना उनसे की जा सकती है, और एक CER/WER पाया जा सकता है। एक आम किताब के लिए कोई गोल्ड मानक लक्षित CER/WER नहीं होता तो शोधकर्ताओं को निष्पादन की गुणवत्ता को ख़ुद ही आंकना होगा और यह फ़ैसला लेना होगा, भाषाई समुदाय

की प्रतिक्रिया के साथ, कि संभावित मॉडलिंग कार्यों में उन्हें कितने CER/WER का लक्ष्य लेकर चलना चाहिए। ध्यान दीजिये कि इसके लिए हमें निष्कर्षित पाठ की एक पंक्ति-संरेखित संस्करण की ज़रूरत पड़ेगी और इसे या तो पंक्ति-स्तरीय ओसीआर से पाया जा सकता है (क्रॉप करने के बाद) या फिर प्रागुक्तित पाठ को संदर्भ पाठ से लेवनश्टैन डिस्टेंस जैसे मापों द्वारा संरेखित कर के।

प्रारंभिक प्रयोग प्रारंभिक प्रयोगों के लिये, हम यह सिफ़ारिश देते हैं कि १-२ मुख्य शोधकर्ता डेटासेट के एक अच्छे ख़ासे प्रतिदर्श को ख़ुद हाथ से अभिटिप्पणित और आडिट करें। यह यह सुनिश्चित करने में मदद कर सकता है कि शोधकर्ता और भाषा समुदाय के सदस्य अभिटिप्पण कार्यप्रवाह से परिचित हैं और आगामी काल के अभिटिप्पणकर्ताओं का सही तरह से मार्गदर्शन कर पायें। ओसीआर प्रयोग चलाने से पहले कुछ अभिटिप्पण कर लेना इसीलिए भी ज़रूरी है क्योंकि अब हम जिन भी माडलों के साथ प्रयोग करेंगे उन सभी के मूल्यांकन के लिए एक मानक समुच्चय होना चाहिए। एक बार जब कुछ पन्ने अभिटिप्पणित हो जायें, तो शोधकर्ता ओसीआर प्रक्रिया को सामान्य ऑफ-द-शेल्फ़ ओसीआर तरीक़ों जैसे गूगल विज़न या ऑक्यूलर द्वारा शुरू कर सकते हैं।

डेटा लेबलीकरण अब जब शोधकर्ताओं को ऑफ-द-शेल्फ़ ओसीआर तरीक़ों की गुणवत्ता का अंदाज़ा हो गया है, हम अगली यह सिफ़ारिश करते हैं कि वह डेटा के एक और भी बड़े प्रतिदर्श का अभिटिप्पण करने के लिए एक अभिटिप्पणकर्ताओं की भर्ती करके, अगर ओसीआर की गुणवत्ता काफ़ी कम पायी गई थी। अभिटिप्पणकर्ताओं को चयनित देशी भाषाओं के देशी वक्ता होने की ज़रूरत नहीं है; पर, उनमें बुनियादी पैटर्न अभिज्ञान, डेटा अभिटिप्पण और टाइपिंग का कौशल होना चाहिए। पूर्व शोधों ने यह दिखाया है कि बिना देशी भाषा के पूर्ण ज्ञान के भी, अभिटिप्पणकर्ता काफ़ी दक्षता से ओसीआर सुधार कार्यों में भाग लेते हैं, यदि वह भाषा की लिपि पढ़ पाएँ एयर यदि वह नये विश्लेषकों में भेद कर पायें (Rijhwani et al., 2023)। अभिटिप्पणकर्ताओं को मानकीकृत दिशानिर्देशों के आधार पर प्रशिक्षित करना चाहिए, और मुख्य शोधकर्ताओं द्वारा एक मैन्युअल ऑडिट ज़रूर किया जाना चाहिए ताकि अनुपालन सुनिश्चित किया जा सके।

पश्चात-सुधार अगर प्रारंभिक प्रयोगों के नतीजे संतोषजनक हैं, तो हम यह सिफ़ारिश करेंगे कि आप इन्हें पश्चात-सुधार के ज़रिये और भी ज़्यादा बेहतर बनाएँ। एक पश्चात-सुधार मॉडल आदर्श रूप से वर्ण-स्तरीय तत्रुटियों को ५% से कम पर ला सकता है (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021) । जैसा कि §3 में परिचर्चित है, हम यह भी सिफ़ारिश करते हैं कि शोधकर्ता प्रतिलिपि प्रणाली, व्याप्ति, विकर्ण/स्थानीय ध्यान, और सक्रिय अधिगम जैसी तकनीकों के संयोजन का इस्तेमाल करें निष्पादन को बेहतर करने के लिए। Rijhwani et al. (2021) अपने शोध में सबसे ज़रूरी पश्चात-सुधार विशेषताओं को कार्यान्वित किया है और उनके कोड के ज़रिए आप सीधे सीधे पश्चात-सुधार माडलों को प्रशिक्षित कर सकते हैं कम-संसाधनीय विन्यासों में।

शुरुआत से प्रशिक्षण दूसरी ओर, अगर प्रारंभिक प्रयोग यह दर्शाते हैं कि त्रुटि दर बहुत अधिक है, तो शोधकर्ता साधारणतः मूल से एक कस्टम ओसीआर प्रणाली को प्रशिक्षित करने पर विचार कर सकते हैं। इसके लिए अच्छी ख़ासी संख्या में प्रशिक्षण के लिए अभिटिप्पणित पन्नों की ज़रूरत पड़ेगी, प्रशिक्षण पाइपलाइन को सेटअप करने में और उसके लिए सबसे श्रेष्ठ हाइपरपरामीटर चुनने में अभिकलनात्मक निपुणता के अलावा। डेटा अभिटिप्पण चरण में एकत्रित की गयीं मानव-अभिटिप्पण को मूल से ओसीआर माडलों को प्रशिक्षित करने के लिए और प्रथम-पारण निर्गमों का इस्तेमाल पश्चात-सुधार माडलों के अधिक प्रशिक्षण के लिए किया जा सकता है। हम कस्टम मॉडल को प्रशिक्षित करने के लिए टेसरेक्ट (Smith, 2007) या ऑक्यूलर (Berg-Kirkpatrick et al., 2013) जैसे ओपन-सोर्स उपकरणों का इस्तेमाल करने की सलाह देते हैं, क्योंकि उनकी दक्षता, इष्टतमीकरण और सक्रिय उपयोगकर्ता समुदाय बेहतर है। अधिक विकसित शोधकर्ता मूल से अपना ख़ुद का आर्किटेक्चर और प्रशिक्षण पाइपलाइन लिखने पर भी विचार कर सकते हैं। हालाँकि, ध्यान दें कि मूल से प्रणाली को प्रशिक्षित करना आसान काम नहीं है, और शोधकर्ताओं को कई चुनौतियों का सामना करना पड़ता है। उदाहरण के लिए, टेसरेक्ट को सेटअप करने में काफ़ी समय लगता है और शुरुआत में बुनियादी चीज़ें सीखना बहुत मुश्किल, इसमें कोई ग्राफ़िकल यूज़र इंटरफ़ेस नहीं है, और इसके लिए उच्च-गुणवत्ता वाली छवियों की आवश्यकता होती है जो हो सकता है कि कुछ कम-संसाधनीय भाषाओं या प्रतिबिंब संग्रहों के लिए उपलब्ध न हों।

तैनाती और सुधार इन-हाउस उपयोग के लिए, अंतिम प्रशिक्षित मॉडल का उपयोग सीधे सीधे पूरे कॉर्पस और उस भाषा में किसी भी अन्य संग्रह को डिजिटलीकृत करने के लिए किया जा सकता है। हम यह सिफ़ारिश करते हैं कि अभिकलनात्मक शोधकर्ता और भाषा समुदाय के सदस्य प्रशिक्षण, अभिटिप्पण और परिनियोजन (डिप्लायमेंट) प्रक्रिया के दौरान संपर्क में रहें और ओसीआर गुणवत्ता और मॉडलिंग में पाए किसी भी मुद्दे को फ़ौरन चिह्नित करें। कुछ मामलों में, यदि उपर्युक्त तकनीकों को आज़माने के बाद भी पर्याप्त ओसीआर गुणवत्ता प्राप्त नहीं हो रही है, तो कुछ रियायतें और आगे डेटा चयन और अभिटिप्पण की आवश्यकता हो सकती है। उदाहरण के लिए, डेटा की गुणवत्ता में सुधार की भी आवश्यकता हो सकती है (जैसे कि मूल स्रोत पाठ की स्कैनिंग को दोबारा करना), अधिक डेटा के लिए अभिटिप्पण का एक और चरण आयोजित करने की आवश्यकता हो सकती है, या विशेष दस्तावेज़ों के लिए अच्छी गुणवत्ता के ओसीआर को प्राप्त करने के लिए कुछ अनूठी एल्गोरिदम तकनीकों को विकसित करने की आवश्यकता भी हो सकती है। हम लेबलस्टूडियो (Tkachenko et al., 2020-2022) का उपयोग करने की सलाह देते हैं, जो एक ओपन-सोर्स लेबलिंग और अभिटिप्पण प्लेटफ़ॉर्म है। उपयोगकर्ता इंटरफ़ेस

उच्च-गुणवत्ता वाला, उपयोगकर्ता के अनुकूल और सहयोगियों और अभिटिप्पणकर्ताओं के साथ सेटअप और साझा करने के लिए काफी सरल है। इसके अलावा एक सक्रिय लेबलस्टूडियो स्लैक भी है जहां समस्याएं अपेक्षाकृत शीघ्रता से हल हो जाती हैं।

6 संबंधित शोध

प्रकाशिक संप्रतीक अबिज्ञान ओसीआर पर एक शोध समस्या के रूप में दशकों से अध्ययन किया गया है, और आज, वाणिज्यिक और ओपन-सोर्स ओसीआर प्रणालियाँ दोनों बखूबी पाठ को अधिकांश प्रतिबिंबों में से सटीक तरह से निष्कर्षित कर सकती हैं, और तो और रियल-टाइम में भी इनका इस्तेमाल किया जा सकता हैं टेस्ट-टाइम दक्षता के कारण (Smith, 2007; Blecher et al., 2023; Berg-Kirkpatrick et al., 2013) । ओसीआर के वर्ण, शब्द, अनुच्छेद का निष्कर्षण शामिल हो सकता है, और किसी पन्ने या प्रतिबिंब के खाके का संरक्षण भी। ओसीआर को डिजिटल मानविकी (Reul et al., 2017; Rijhwani et al., 2021, 2020) और व्यापार में काफ़ी इस्तेमाल किया जाता है क्योंकि विरल पाण्डुलिपियों, किताबों, भाषावैज्ञानिक फील्ड नोट्स, इनवॉइस, वाणिज्यिक दस्तावेज़, इत्यादि के डिजिटलीकरण के लिये यह एक आवश्यक चरण है। अनुप्रवाही एनएलपी कार्यों और अनुप्रयोगों के लिए कम-संसाधनीय भाषाओं में डेटा सृजन के लिए भी यह एक बहुमूल्य तकनीक है (Ignat et al., 2022b)। पिछले दो दशकों में, कंप्यूटर दृष्टि शोध समुदाय ने ओसीआर घटनाक्रमों को कवर करते हुए कई श्रेष्टतम सर्वेक्षण प्रकाशित करे हैं (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020) । कम-संसाधनीय प्रक्षेत्र में, Hedderich et al. (2021) का सर्वेक्षण व्यापक एनएलपी प्रगतियों को कवर करता है पर प्रकाशिक संप्रतीक अभिज्ञान को नहीं। हमारी जानकारी के अनुसार, हमारे अलावा किसी अन्य शोधकार्य ने कम-संसाधनीय भाषाओं के लिए ओसीआर को सर्वेक्षित नहीं किया है।

संसाधन सृजन पाठ या प्रतिबिंब आधारित डेटासेटों और कॉर्पोरा आम तौर पर वेब को खुरच (स्क्रेप) और रेंग (क्रॉल) कर निर्मित किए जाते हैं; लेकिन, हम कुछ ऐसे ओसीआर-निर्मित डेटासेटों पर प्रकाश डालना चाहते हैं, ख़ास तौर से वह जो टेबल 1 में मौजूद भाषाओं के अतिरिक्त महामेरिकी देशी भाषाओं के साथ काम करते हैं । Cordova and Nouvel (2021) मध्य केचुआ में संसाधनों की कमी को संबोधित करते हैं, क्योंकि संसाधन प्रायः प्रभुत्वशाली दक्षिणी क़िस्म में ही होते हैं, और ओसीआर प्रोद्योगिकिओं के ज़रिये एक सफलतापूर्वक डिजिटलीकृत कार्पस को साझा करते हैं। Hunt et al. (2023) एक अकूज़िपिक (अलास्का और रुस के कुछ हिस्सों में बोले जाने वाली एक देशी भाषा) शब्दकोश को साझा करते हैं जिसमें समांतर रूसी भाषी पाठ निहित हैं, जो अनुप्रवाही एनएलपी कार्यों के लिए बहुत फ़ायदेमंद पाया गया है। कुछ अन्य प्रासंगिक लेकिन ग़ैर-ओसीआर डेटासेट सृजन प्रयासों में शामिल हैं गुआरानी-स्पेनी न्यूज़ आर्टिकल (Góngora et al., 2021) , नाहुआत्ल वाणी अनुवाद (Shi et al., 2021), और मज़ातेक और मिश्तेक अनुवाद (Tonja et al., 2023) , जो सभी ओसीआर के लिये पूर्वप्रशिक्षण कॉर्पोरा के रूप में बहुत की लाभदायक हो सकते हैं।

7 निष्कर्ष

इस पेपर में, हमने उन प्रकाशिक सम्प्रतीक अभिज्ञान (ओसीआर) तकनीकों का संक्षिप्त सर्वेक्षण प्रस्तुत किया है जो ओसीआर साहित्य में कम-संसाधनीय भाषाओं के लिये एप्लीकेबल* साबित हुई हैं। यह सर्वेक्षण फोकस्ड* है और ऐसा काम पहले प्रकाशित नहीं हुया है चूँकि कम-संसाधनीय और देशी भाषाओं के ओसीआर में काम करने वाले शोधकर्ताओं का समुदाय आज भी छोटा ही माना जाता है। हम इस पेपर में १० केंद्रीय और दक्षिण अमेरिकी देशी भाषाओं के कुछ एडिजिटलीकृत डेटासेटों पर प्रकाश डालते हैं, जो कि मुख्यतः आईला संग्रह से हैं, जिन्हें अनुप्रवाह एनएलपी एप्लिकेशनों* में इस्तेमाल के लिए डिजिटलीकृत करना बहुत फ़ायदेमंद साबित होगा। हमारे एक्सपीरियंस* और हमारे लिटरेचर-रिव्यू* के नतीजों के आधार पर, हम अभिकलनात्मक शोधकर्ताओं, भाषावैज्ञानिकों, और भाषा समुदायों के फ़ायदे के लिए ओसीआर को बेहतर करने और उसे इस्तेमाल करने के उद्देश्य के लिए कुछ सिफ़ारिशों से अपने पेपर को निष्कर्षित करते हैं। हम यह आशा करते हैं कि इसे एक स्टार्टिंग पॉइंट* की तरह शोधकर्ता या भाषा समुदाय इस्तेमाल करेंगे, ख़ास कर के वह जो अपनी भाषा के संसाधनों और दस्तावेज़ों को डिजिटलीकृत करना चाहते हैं पर यह नहीं जानते कि कौनसी ओसीआर एडाप्टेशंस* आज के ज़माने में बेहद ज़रूरी बन चुकी हैं एक श्रेष्ठ-गुणवत्ता का ओसीआर आउटपुट प्राप्त करने के लिए और इस क्षेत्र में अभी भी क्या क्या खुली चुनौतियाँ हैं।

सीमाएँ

हम यह एकनॉलेज* करते हैं कि इस लौंग-पेपर की पृष्ठ-सीमा के अंदर, हमारे विषय की सभी पेचीदगियों को शामिल करना मुमकिन नहीं है। जहाँ हो सके, हमने रिलेवेंट* डेटेल्स* को पेश किया है, जैसे गणितीय समीकरण, फिगर*, और टेबल*, ताकि सर्वेक्षण को संक्षिप्त और अमेरिकाज-एनएलपी समुदाय के लिए रिलेवेंट* रखा जा सके। इसके साथ ही, सर्वेक्षण को रूपरेखा और उद्देश्य देती १० देशी भाषाओं में एक्सपेरिमेंटल* नतीजे इस पेपर के स्कोप* से बाहर हैं और आसानी से अलग पेपर बन सकते हैं।

नैतिक वक्तव्य

१० चयनित देशी भाषाओं के रॉ* डेटा संसाधन पूर्णतः आईला द्वारा होस्टेड* हैं। डेटा आम जनता के लिए मुफ़्त रूप से उपलब्ध हैं, और कुछ फ़ाइलें* दरखास्त* द्वारा शेयर* की जा सकती हैं। ओवरल*, डेटा को बिना अनुमति लिए इस्तेमाल किया जा सकता है, और बिना किसी फ़ीस* चुकाए, जब तक संसाधन को सही तरीक़े से उद्धृत किया जाये। हम उन भाषावैज्ञानिकों, नेटिव* और हेरिटेज* भाषा-भाषियों, और आईला टीम के शुक्रगुज़ार हैं जिन्होंने लातिनी

अमेरिका की देशी भाषाओं के लिये इतनी मूल्यवान रिपॉज़िटरी* की रचना करी। इस काम की एक नैतिक इम्प्लिकेशन* यह है कि यह भाषा संसाधन निर्माण और प्राकृतिक भाषा प्रसंस्करण में ज़्यादा से ज़्यादा सस्टेनेबल* और एक्विटेबल* कार्यों को प्रोत्साहित करेगा। लेकिन हम इस काम से, जो ओसीआर माडलों की ओपन-सोर्स डेवलपमेंट* को प्रोत्साहन देता है ताकि शोधकर्ता कमेरिशल* सिस्टमों से दूर हैट सकें कम-संसाधनीय और देशी भाषा डेटा को प्रसंस्कृत करने के लिए, उत्पन्न होते कोई भी नकारात्मक नैतिक चिंताएँ नहीं देखते।

आभार

इस परियोजना को National Endowment for the Humanities द्वारा अवार्ड संख्या PR-276810-21 के अन्तर्गत जेनरस* वित्त-पोषण मिला है। लेखक बेनामी रिव्यूअरों* के वैल्युएबल* सिफ़ारिशों, फीडबैक*, और कमेंटों* के भी आभारी हैं।

उद्धरण

Ahmad Abdulkader and Mathew R. Casey. 2009. Low cost correction of ocr errors using learning in a multi-engine environment. In *2009 10th International Conference on Document Analysis and Recognition*, pages 576–580.

Eric Adell, Antonio Moisés Toma Cruz, and Edelmira Catarina Sánchez Toma. 2016. Kam nib'anax tu ma'l xhemaana (a brief description of a typical week). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119533. Accessed March 21, 2024. IXIL-CTZ-DES-EST-2016-06-23-0507.

Reem Alaasam, Berat Kurar, and Jihad El-Sana. 2019. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 738–742. IEEE.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. In *Linguistic Issues in Language Technology*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 207–217. The Association for Computer Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 118–123, Baltimore, Maryland. Association for Computational Linguistics.

Tulio Bermúdez Mejía. 2015. Miskitu dance, food, and traditions: traditional miskitu food, dance, songs, festivities. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119700. Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents.

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 87–107, Singapore. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of Ancash lexical resources from OCR. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 163–167, Online. Association for Computational Linguistics.

Hervé Déjean and Jean-Luc Meunier. 2019. Table rows segmentation. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 461–466. IEEE.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2017. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1006–1014, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. OCR and post-correction of historical Finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76, Gothenburg, Sweden. Association for Computational Linguistics.

Pierre Déléage. 2002. Sharanahua language collection of pierre déléage. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Nora England. 1972-1985. Mam language stories and grammars. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119520, ailla:119520,ailla:119520,ailla:119520,ailla:119520. Accessed February 15, 2024.

Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C. Popat. 2017. Sequence-to-label script identification for multilingual OCR. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 161–168. IEEE.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 467–472, San Diego, California. Association for Computational Linguistics.

Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Anshul Gupta, Ricardo Gutierrez-Osuna, Matthew Christy, Richard Furuta, and Laura Mandell. 2016. Font identification in historical documents using active learning. *CoRR*, abs/1601.07252.

Harsh Gupta, Luciano Del Corro, Samuel Broscheit, Johannes Hoffart, and Eliot Brenner. 2021. Unsupervised multi-view post-OCR error correction with language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8647–8652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Nicholas Hopkins. 1964. A dictionary of the chuj (mayan) language language community. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119647. Accessed February 15, 2024.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022a. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022b. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

José Carlos Aradillas Jaramillo, Juan José Murillo-Fuentes, and Pablo M. Olmos. 2018. Boosting handwriting text recognition in small databases with transfer learning. In *16th International Conference on*

Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018, pages 429–434. IEEE Computer Society.

Zhaohui Jiang, Zheng Huang, Yunrui Lian, Jie Guo, and Weidong Qiu. 2019. Integrating coordinates with context for information extraction in document images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 363–368. IEEE.

Heidi Anna Johnson. 2000-2005. A grammar of san miguel chimalapa zoque. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119500 Accessed February 15, 2024.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Susan Kalt. 2016. Entrevista con tomas castro v y santusa quispe de flores. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Terrence Kaufman. 1960-1993. Colección de idiomas mayenses de terrence kaufman. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024.

Umar Khan, Sohaib Zahid, Muhammad Asad Ali, Adnan Ul-Hasan, and Faisal Shafait. 2021. Tabaug: Data driven augmentation for enhanced table structure recognition. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 585–601. Springer.

Shachar Klaiman and Marius Lehne. 2021. Docreader: Bounding-box free training of a document information extraction model. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 451–465. Springer.

Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 867–874. The Association for Computational Linguistics.

Amrith Krishna, Bodhisattwa P. Majumder, Rajesh Bhat, and Pawan Goyal. 2018. Upcycle your OCR: Reusing OCRs for post-OCR text correction in Romanised Sanskrit. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 345–355, Brussels, Belgium. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13094–13102. AAAI Press.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Manfei Liu, Zecheng Xie, Yaoxiong Huang, Lianwen Jin, and Weiyin Zhou. 2018. Distilling gru with data augmentation for unconstrained handwritten text recognition. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 56–61.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *CoRR*, abs/2206.00311.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu. Association for Computational Linguistics.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access*, 8:142642–142668.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960, Austin, Texas. Association for Computational Linguistics.

Claire Monteleoni and Matti Kaariainen. 2007. Practical online active learning for classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Marcin Namysl and Iuliu Konya. 2019. Efficient, lexicon-free ocr using deep learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 295–301.

Andrew Naoum, Joel Nothman, and James R. Curran. 2019. Article segmentation in digitised newspapers with a 2d markov model. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1007–1014. IEEE.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, and Andy Way. 2020. A tool for facilitating OCR postediting in historical documents. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 47–51, Marseille, France. European Language Resources Association (ELRA).

Animesh Prasad, Hervé Déjean, and Jean-Luc Meunier. 2019. Versatile layout understanding via conjugate graph. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 287–294. IEEE.

José Ramón Prieto and Enrique Vidal. 2021. Improved graph methods for table layout understanding. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 507–522. Springer.

Paul Proulx. 1968. Chiquian quechua vocabulary. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Abhishek Prusty, Sowmya Aitha, Abhishek Trivedi, and Ravi Kiran Sarvadevabhatla. 2019. Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 999–1006. IEEE.

Christian Reul, Uwe Springmann, and Frank Puppe. 2017. LAREX - A semi-automatic open-source tool for layout analysis and region extraction on early printed books. *CoRR*, abs/1701.07396.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR accuracy on early printed books by combining pretraining, voting, and active learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime G. Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6924–6931. AAAI Press.

Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1703–1714, Osaka, Japan. The COLING 2016 Organizing Committee.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Burr Settles. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Prema Satish Sharan, Sowmya Aitha, Amandeep Kumar, Abhishek Trivedi, Aaron Augustine, and Ravi Kiran Sarvadevabhatla. 2021. Palmira: A deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 477–491. Springer.

Zejiang Shen, Weining Li, Jian Zhao, Yaoliang Yu, and Melissa Dell. 2022. OLALA: Object-level active learning for efficient document layout annotation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 170–182, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.

Nikita Srivatsan, Si Wu, Jonathan Barron, and Taylor Berg-Kirkpatrick. 2021. Scalable font reconstruction with dual latent manifolds. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3060–3072, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Storchan and Jocelyn Beauschene. 2019. Data augmentation via adversarial networks for optical character recognition/conference submissions. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 184–189.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ahmad Pahlavan Tafti, Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy L. Peissig. 2016. OCR as a service: An experimental evaluation of google docs ocr, tesseract, ABBYY finereader, and transym. In *Advances in Visual Computing - 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I*, volume 10072 of *Lecture Notes in Computer Science*, pages 735–746. Springer.

Lindia Tjuatja, Shruti Rijhwani, and Graham Neubig. 2021. Explorations in transfer learning for ocr post-correction. In *Fifth Widening Natural Language Processing Workshop (WiNLP)*, volume 6.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for historical corpora: An assessment on post-ocr correction and named entity recognition. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*, volume 2723 of *CEUR Workshop Proceedings*, pages 310–339. CEUR-WS.org.

Atnafu Lambebo Tonja, Christian Maldonado-sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (Americas-NLP)*, pages 94–102, Toronto, Canada. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Nikolai Vogler, Jonathan Allen, Matthew Miller, and Taylor Berg-Kirkpatrick. 2022. Lacuna reconstruction: Self-supervised pre-training for low-resource historical document transcription. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 206–216, Seattle, United States. Association for Computational Linguistics.

Robin M. Wright, Manuel da Silva, and José Felipe Aguiar. 2000. Baniwa history: Uapui cachoeira, aiary river (1970s - 2000). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119657 Accessed March 21, 2024.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy parallel data alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Dubrovnik, Croatia. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.