E-ISSN-2584-2870 DOI: https://doi.org/10.65833/dbs.2026.011

देवभूमि-सौरभम्
Devabhūmisaurabham

Vol 2 Issue 2 July-Dec 2025 Devabhūmisaurabham International Journal of CSU Shri Raghunath Kirti Campus Devprayag Uttarakhand India

कम्प्यूटर युग में भारतीय भाषा विज्ञान: स्वर्णिम अवसर एवं जटिल चुनौतियाँ

लेखक - सिद्धार्थ तिवारी
शोधच्छात्र
संस्था - संस्कृत विभाग, कला संकाय, काशी हिन्दू विश्वविद्यालय
ई-मेल - tiwarisiddharth@bhu.ac.in
Mo. No. – 7497990957
सह-लेखक – अभिनेष कुमार मिश्र
शोधच्छात्र
संस्था - संस्कृत विभाग, कला संकाय, काशी हिन्दू विश्वविद्यालय
Email – akmishra999@bhu.ac.in
Mo. No. - 9648798881

सार

इक्कीसवीं सदी के तृतीय दशक में, जब विश्व Generative AI और "लार्ज लैंग्वेज मॉडल्स" (LLMs) की क्रान्ति का साक्षी बन रहा है, भारतीय भाषा विज्ञान एक ऐतिहासिक मोड़ पर खड़ा है। इस शोधपत्र का उद्देश्य २१वीं सदी के डिजिटल एवं कम्प्यूटर युग में भारतीय भाषा विज्ञान के समक्ष उपस्थित गतिशील परिवर्तनों का गहन विश्लेषण प्रस्तुत करना है। कम्प्यूटेशनल लिंग्विस्टिक्स, प्राकृतिक भाषा प्रसंस्करण (NLP) एवं आर्टिफिशियल इंटेलिजेंस जैसे क्षेत्रों में हो रही प्रगति ने भारतीय भाषाओं के अध्ययन एवं अनुप्रयोग के लिए अभूतपूर्व सम्भावनाएँ उत्पन्न की हैं। विशेषतः संस्कृत भाषा के पाणिनीय व्याकरण जैसी अत्यन्त वैज्ञानिक एवं नियम-आधारित परम्पराएँ आधुनिक भाषा प्रौद्योगिकी के विकास में एक मौलिक आधार प्रदान कर सकती हैं। परन्तु साथ ही, इन अवसरों का पूर्ण लाभ उठाने के मार्ग में गम्भीर चुनौतियाँ भी विद्यमान हैं; जैसे डिजिटल संसाधनों (एनोटेटेड कोष - जहाँ प्रत्येक शब्द के अर्थ के साथ अतिरिक्त जानकारी, जैसे व्याकरणिक विवरण या सांस्कृतिक सन्दर्भ दिए जाते हैं, ताकि पाठक को गहरी समझ मिल सके) का अभाव, तकनीकी मानकीकरण की कमी, बहुभाषिक अन्तरापृष्ठ का अविकसित होना, तथा शैक्षणिक शोध एवं औद्योगिक आवश्यकताओं के बीच का अन्तराल। यह शोधपत्र इन समस्त पहलुओं का विस्तृत परीक्षण करते हुए एक सुदृढ़ रोडमैप प्रस्तावित करता है, जिसमें अन्तःविषयक शिक्षा को बढ़ावा, राष्ट्रीय स्तर पर भाषा प्रौद्योगिकी मिशन की स्थापना, तथा खुले स्रोत के संसाधनों के सहयोगात्मक निर्माण जैसे प्रमुख सुझाव सम्मिलित हैं। निष्कर्षतः, यह शोधपत्र इस तथ्य की पुष्टि करता है कि यदि इन चुनौतियों का समाधान एक समग्र रणनीति के अन्तर्गत किया जाए, तो भारतीय भाषा विज्ञान न केवल अपनी ऐतिहासिक गरिमा को पुनः प्राप्त कर सकता है, बल्कि वैश्विक डिजिटल समावेशन एवं बहुभाषिक कम्प्यूटिंग के क्षेत्र में एक अग्रणी भूमिका भी निभा सकता है।

कूटशब्द - भारतीय भाषा विज्ञान, कम्प्यूटेशनल लिंग्विस्टिक्स, प्राकृतिक भाषा प्रसंस्करण (NLP), पाणिनीय व्याकरण, बहुभाषिक प्रौद्योगिकी, डिजिटल संसाधन, भाषाई डेटा सेट, आर्टिफिशियल इंटेलिजेंस, भारतीय भाषाओं का संरक्षण।

प्रस्तावना: एक परिवर्तनशील परिदृश्य

भारतीय उपमहाद्वीप न केवल भाषाई विविधता का केन्द्र रहा है, बल्कि भाषा-विज्ञान की जन्मस्थली भी है। लगभग २५०० वर्ष पूर्व महर्षि पाणिनि ने 'अष्टाध्यायी' के रूप में जिस वैज्ञानिक व्याकरण की नींव रखी थी, आज वह २१वीं सदी के कम्प्यूटर युग में सबसे अधिक प्रासंगिक हो गई है। वर्तमान वर्ष २०२५ में, जब इण्टरनेट

और कृत्रिम बुद्धिमत्ता मानव जीवन का अभिन्न अङ्ग बन चुके हैं, भाषाओं का अस्तित्व केवल मौखिक संवाद तक सीमित नहीं रह गया है, बल्कि वह "डेटा" और "एल्गोरिद्म" का रूप ले चुकी हैं।

आज ChatGPT, Gemini और Llama जैसे "लार्ज लैंग्वेज मॉडल्स" (LLMs) ने सूचना तकनीक की दुनिया बदल दी है। परन्तु, विडम्बना यह है कि विश्व की सबसे पुरानी और वैज्ञानिक भाषा परम्परा होने के बावजूद, भारतीय भाषाएँ डिजिटल स्पेस में संसाधनों की कमी से जूझ रही हैं। पश्चिमी AI मॉडल्स अक्सर भारतीय भाषाओं के सन्दर्भ, संस्कृति और व्याकरण को समझने में त्रुटियाँ करते हैं, जिसे तकनीकी भाषा में Hallucination कहा जाता है। पाणिनि द्वारा रचित "अष्टाध्यायी" न केवल संस्कृत का, बल्कि किसी भी मानव भाषा का प्रथम पूर्णतः वैज्ञानिक एवं नियमबद्ध व्याकरण माना जाता है पाणिनि का व्याकरण वैज्ञानिक, नियमबद्ध एवं एल्गोरिदमिक है क्योंकि यह भाषा को ध्वन्यात्मक, रूपात्मक एवं वाक्य-संरचनात्मक स्तरों पर पूर्णतः व्यवस्थित नियमों द्वारा वर्णित करता है, जिसमें कोई अपवाद या अस्पष्टता नहीं रह जाती।

वैज्ञानिकता - पाणिनि ने भाषा को ध्वनियों, शब्द-व्युत्पत्ति एवं वाक्य-विन्यास के आधार पर वर्गीकृत किया, जिसमें **माहेश्वर सूत्र**[1] ध्वनियों का वैज्ञानिक क्रम स्थापित करते हैं तथा सञ्ज्ञा सूत्र रूप, कारक आदि की परिभाषाएँ निर्धारित करते हैं। यह प्रणाली आधुनिक भाषाविज्ञान की भाँति वर्णनात्मक एवं उत्पादक है, जहाँ 4000 सूत्रों से समस्त संस्कृत रूप-रचना युक्तिसङ्गत रूप से व्युत्पन्न होती है।

नियमबद्धता - अष्टाध्यायी में प्रत्येक नियम परस्परावलम्बी एवं क्रमबद्ध है, जहाँ अनुवृत्ति (पूर्व नियमों का अनुसरण) तथा अधिकार सूत्र जो स्वयं कोई कार्य (क्रिया) नहीं बताता, बल्कि अपने बाद आने वाले कई सूत्रों के लिए एक सन्दर्भ या विषय-क्षेत्र निर्धारित करता है, जिससे उन सूत्रों को बार-बार वह विषय दोहराना न पड़ता। एवं नियम-संघर्ष-समाधान जैसे **विप्रतिषेधे परं कार्यम्**[2] जब दो नियम एक साथ लागू होते हैं, तो बाद वाला कार्य पहले वाले पर श्रेष्ठता पाता है, यह परिभाषा सूत्र होने के कारण नियमों के टकराव को सुलझाता है। सूत्रों की प्रतिष्णात शैली (अक्षरसङ्क्षिप्तता) एवं सहायक प्रतीक जो प्रत्याहार के माध्यम से जैसे इक: , यण् अचि आदि) भाषा-उत्पादन को कठोर अनुशासन प्रदान करते हैं, जिससे अपाणिनीय प्रयोग अस्वीकार्य हो जाते हैं।

एल्गोरिदमिक संरचना - यह एक उत्पादक व्याकरण (generative grammar) है, जहाँ धातु+प्रत्यय से शब्दरूप क्रमिक नियमों द्वारा स्वचालित रूप से निर्मित होते हैं, ठीक वैसे ही जैसे आधुनिक कम्प्यूटर प्रोग्रामिंग में **सिन्टैक्स नियम**[3] कार्य करते हैं। **"विप्रतिषेधे परं कार्यम्"** सूत्र के अनुसार, पाणिनीय नियम पूर्व सूत्र की अपेक्षा परसूत्र में लागू होते हैं, जिससे मशीन द्वारा लाखों शब्द अपवादरहित रूप से उत्पन्न हो सकते हैं। इस कारण पाणिनीय तन्त्र को विश्व का प्रथम औपचारिक भाषा-एल्गोरिदम माना जाता है। इसी प्रकार द्रविड़ भाषा परिवार में 'तोल्काप्पियम्' (तमिल) तथा कर्नाटक में 'कविराजमार्ग' (कन्नड़) जैसे ग्रन्थों ने भाषा के काव्यशास्त्रीय एवं व्याकरणिक सिद्धान्तों को स्थापित किया।

२०वीं सदी के उत्तरार्ध में कम्प्यूटर के उदय एवं २१वीं सदी में इण्टरनेट तथा डिजिटल प्रौद्योगिकी के विस्फोटक विकास ने मानव ज्ञान के प्रत्येक क्षेत्र को परिवर्तित कर दिया है। भाषा विज्ञान भी इससे अछूता नहीं रहा। **कम्प्यूटेशनल लिंग्विस्टिक्स**[4] नामक एक नवीन अन्तःविषयक शाखा का उदय हुआ, जिसका उद्देश्य कम्प्यूटेशनल मॉडलों एवं एल्गोरिदम्स के माध्यम से भाषा की संरचना, प्रयोग एवं अर्जन को समझना है। इसी का व्यावहारिक अनुप्रयोग है **प्राकृतिक भाषा प्रसंस्करण**[5] (NLP), जो मशीनों को मानव भाषा को समझने, विश्लेषण करने एवं उत्पन्न करने की क्षमता प्रदान करता है।

इस नए युग में भारतीय भाषा विज्ञान की भूमिका द्वि-आयामी है -

---

[1] अइउण् ऋलृक् एओङ् ऐऔच् हयवरट् लण् ञमङणनम् झभञ् घढधष् जबगडदश् खफछठथचटतव् कपय् शषसर् हल्। इति माहेश्वराणि सूत्राणि।

[2] पाणिनीय अष्टाध्यायी १/४/२

[3] सिंटैक्स नियम वे होते हैं जो बताते हैं कि शब्दों, चिह्नों या तत्वों को एक साथ कैसे व्यवस्थित किया जाए ताकि एक सही और अर्थपूर्ण वाक्य, कोड या संरचना बन सके; इसमें शब्दों का क्रम (जैसे हिन्दी में कर्ता-कर्म-क्रिया), क्रिया-कर्ता का मेल, विराम चिह्नों का सही उपयोग और कम्प्यूटर भाषाओं में विशेष प्रतीकों (जैसे {} या ;) का सही प्रयोग शामिल है।

[4] कम्प्यूटेशनल लिंग्विस्टिक्स एक अन्तःविषय क्षेत्र है जो भाषा और कंप्यूटर विज्ञान को जोड़ता है, जिसका उद्देश्य कंप्यूटर को मानव भाषा को समझना, व्याख्या करना और उत्पन्न करना सिखाना है; यह प्राकृतिक भाषा प्रसंस्करण (NLP), मशीन लर्निंग, और कृत्रिम बुद्धिमत्ता का उपयोग करके भाषा के गणितीय मॉडल बनाता है और इसमें स्पीच रिकग्निशन, मशीन ट्रांसलेशन, और टेक्स्ट एनालिसिस जैसे एप्लीकेशन शामिल हैं।

[5] प्राकृतिक भाषा प्रसंस्करण (NLP) AI की एक शाखा है जो कंप्यूटर को मानव भाषा (लिखी और बोली जाने वाली) को समझने, व्याख्या करने और उत्पन्न करने में सक्षम बनाती है, जिससे मशीनें इंसानों की तरह भाषा के साथ बातचीत कर सकें; यह टेक्स्ट को प्रोसेस करके डेटा से जानकारी निकालने, भावनाओं को समझने और मशीनों को मानवीय भाषा में प्रतिक्रिया देने में मदद करता है, और इसके उपयोग चैटबॉट, स्पीच रिकॉग्निशन और अनुवाद जैसे क्षेत्रों में होते हैं।

१.एक उपभोक्ता के रूप में - विश्व की दूसरी सबसे बड़ी जनसंख्या द्वारा बोली जाने वाली विविध भाषाओं (संस्कृत, हिन्दी, बांग्ला, तेलुगु, मराठी, तमिल आदि) के रूप में, इनके लिए डिजिटल साधनों (जैसे खोज इंजन, अनुवादक, आभासी सहायक) का विकास एक विशाल सामाजिक तथा आर्थिक आवश्यकता है।

२.एक योगदानकर्ता के रूप में - अपनी अत्यन्त सुव्यवस्थित एवं नियमबद्ध व्याकरणिक परम्पराओं के कारण, भारतीय भाषाएँ विशेषकर संस्कृत, आधुनिक NLP एवं AI मॉडलों के विकास के लिए एक सैद्धान्तिक एवं प्रायोगिक कोष प्रस्तुत करती हैं।

यह शोधपत्र इसी द्वन्द्व एवं समन्वय के बीच के परिदृश्य का अवलोकन प्रस्तुत करेगा।

अवसरों का स्वर्णिम क्षेत्र

प्राचीन सिद्धान्तों पर आधारित आधुनिक एल्गोरिदम - पाणिनीय व्याकरण की मौलिक प्रकृति इसे कम्प्यूटेशन के लिए अद्वितीय रूप से उपयुक्त बनाती है।

सीमित अवस्था मशीनों (Finite State Machines) के साथ साम्य - संस्कृत की शब्द-रचना अत्यन्त समृद्ध एवं नियमबद्ध है। प्रत्ययों (प्रत्यय, विभक्तियाँ) के योग से शब्दों का निर्माण होता है। इस प्रक्रिया को फाइनाइट स्टेट ट्रांसड्यूसर (FST) नामक कम्प्यूटेशनल मॉडल के द्वारा अत्यन्त कुशलता से प्रदर्शित एवं कार्यान्वित किया जा सकता है । FST एक सीमित संख्या में अवस्थाओं एवं संक्रमणों वाली मशीन है, जो एक इनपुट स्ट्रिंग को नियमों के अनुसार प्रसंस्कृत कर एक आउटपुट स्ट्रिंग में बदल सकती है। उदाहरण के लिए, धातु 'पठ्' से 'पठति', 'पठन्ति', 'पठिष्यति' आदि रूपों का निर्माण FST द्वारा मॉडल किया जा सकता है। इस तकनीक का प्रयोग इनफ्लेक्सनल मॉर्फोलॉजी के विश्लेषण एवं संश्लेषण के लिए किया जाता है।

ज्ञान-प्रतिनिधित्व एवं ऑन्टोलॉजी - संस्कृत के कोश (थिसॉरस) जैसे 'अमरकोश' मात्र शब्दकोश नहीं हैं, बल्कि शब्दों के बीच सम्बन्धों (पर्यायवाची, विपरीतार्थक, हाइपोनीमी-हाइपरनीमी) का एक जाल प्रस्तुत करते हैं। यह संकल्पना आधुनिक 'वर्डनेट' (एक शब्दार्थिक सम्बन्ध नेटवर्क) एवं ऑन्टोलॉजी (अवधारणाओं एवं उनके अन्तर्सम्बन्धों का औपचारिक विवरण) के सिद्धान्तों से सीधे मेल खाती है। **इण्डोवर्डनेट परियोजना**[6] इसी सिद्धान्त पर कार्य करती हुई विभिन्न भारतीय भाषाओं के वर्डनेट्स को आपस में जोड़ती है। यह बहुभाषिक शब्दार्थिक खोज एवं अर्थ-आधारित अनुवाद के लिए आधारभूत संरचना प्रदान करता है।

शब्द-विच्छेद में अग्रणी भूमिका - अधिकांश भारतीय भाषाएँ "संश्लेषणात्मक" हैं, अर्थात एक शब्द में कई व्याकरणिक जानकारियाँ (लिङ्ग, वचन, कारक, काल) समाहित होती हैं। इसलिए किसी भी उच्च-स्तरीय प्रसंस्करण (जैसे वाक्य-विश्लेषण, अनुवाद) से पहले शब्द का उसके मूल (धातु/प्रातिपदिक) एवं प्रत्ययों में विभाजन अनिवार्य है। संस्कृत व्याकरण में इस "विग्रह" की अवधारणा पहले से ही सुस्पष्ट है। इस सिद्धान्त पर आधारित टूलकिट्स (जैसे **Samsadhani, Sanskrit Computational Toolkit**[7]) का विकास हुआ है, जिनके सिद्धान्तों को अन्य भारतीय भाषाओं के लिए अनुकूलित किया जा सकता है।

बहुभाषिक भारत के लिए प्रौद्योगिकी का निर्माण

भारत की भाषाई विविधता एक चुनौती होने के साथ-साथ एक विशाल बाजार एवं अनुसन्धान का क्षेत्र भी है।

स्वचालित अनुवाद (Machine Translation - MT) - शिक्षा, प्रशासन, न्यायपालिका एवं व्यापार में भाषा की बाधा को दूर करने के लिए उच्च-गुणवत्ता वाले MT सिस्टम्स की आवश्यकता है। गूगल ट्रांसलेट, माइक्रोसॉफ्ट ट्रांसलेटर जैसे उपकरणों में भारतीय भाषाओं का समर्थन बढ़ा है, परन्तु

[6] इण्डोवर्डनेट भारत की 18 अनुसूचित भाषाओं , जैसे असमिया, बांग्ला, बोडो, गुजराती, हिन्दी, कन्नड़, कश्मीरी, कोंकणी, मलयालम, मणिपुरी, मराठी, नेपाली, ओडिया, पंजाबी, संस्कृत, तमिल, तेलुगु और उर्दू के शब्द जालों का एक संबद्ध शाब्दिक ज्ञान आधार है। शब्द जाल प्राकृतिक भाषा प्रसंस्करण, सूचना निष्कर्षण, शब्द अर्थ स्पष्टीकरण और पाठ से संबंधित अन्य प्रकार के कम्प्यूटेशनल प्रसंस्करण के लिए उपयोग किए जाने वाले शब्दों के कम्प्यूटेशनल डेटाबेस हैं।

[7] संस्कृत कम्प्यूटेशनल टूलकिट में वे सभी डिजिटल उपकरण और सॉफ्टवेयर शामिल हैं जो संस्कृत भाषा के विश्लेषण, प्रसंस्करण और शिक्षण में मदद करते हैं, जैसे "संस्कारधानी" जो संधि-विच्छेद, व्याकरण विश्लेषण, और पद्य को गद्य में बदलने जैसे कार्य करता है, साथ ही Morphological Generators और टेक्स्ट-टू-स्पीच इंजन भी, जो संस्कृत को समझना और सिखाना आसान बनाते हैं।

विशेषकर लघु भाषाओं के लिए एवं साहित्यिक/तकनीकी पाठ के लिए इनकी गुणवत्ता सीमित है। नियम आधारित, सांख्यिकीय एवं **न्यूरल मशीन अनुवाद**[8] जैसी विधियों पर शोध की अपार सम्भावना है।

सूचना पुनर्प्राप्ति एवं खोज इंजन - हिन्दी, तमिल आदि में वेब सामग्री की मात्रा में विस्फोट हुआ है। इन भाषाओं के लिए ऐसे खोज इंजनों की आवश्यकता है जो न केवल शब्दों का मिलान करें, बल्कि समानार्थी शब्दों, रूपभेदों एवं प्रसङ्ग को समझ सकें। यहाँ इण्डोवर्डनेट जैसे संसाधन महत्वपूर्ण हो जाते हैं।

वाक् प्रौद्योगिकी - भारत में अर्धशहरी एवं ग्रामीण क्षेत्रों में तथा कम साक्षरता वाले समुदायों के लिए, आवाज-से-पाठ (Speech-to-Text) एवं पाठ-से-आवाज (Text-to-Speech) प्रणालियाँ डिजिटल समावेशन की कुंजी हैं। इन प्रणालियों के विकास के लिए भाषा-विशिष्ट ध्वनि संग्रह (Speech Corpora[9]) एवं ध्वनि मॉडल (Acoustic Models[10]) के निर्माण की आवश्यकता है, जो भाषाविज्ञान एवं कम्प्यूटर विज्ञान के संयुक्त प्रयास से ही सम्भव है।

डिजिटल मानविकी एवं संरक्षण

प्राचीन पाण्डुलिपियों का डिजिटलीकरण एवं विश्लेषण - **ऑप्टिकल कैरेक्टर रिकग्निशन**[11] (OCR) तकनीक के विकास से संस्कृत, प्राकृत, अपभ्रंश आदि की लाखों पाण्डुलिपियों को डिजिटल रूप दिया जा सकता है। इसके बाद **टेक्स्ट माइनिंग**[12] के द्वारा इन विशाल कोषों (Corpora) में छुपे हुए विचारों, शैलियों, ऐतिहासिक परिवर्तनों आदि का पता लगाया जा सकता है। यह साहित्यिक शोध को क्रान्तिकारी रूप से बदल सकता है।

लुप्तप्राय भाषाओं का दस्तावेजीकरण - भारत में सैकड़ों लघु एवं आदिवासी भाषाएँ लुप्त होने की दशा में हैं। इन भाषाओं के ऑडियो/वीडियो रिकॉर्डिंग, शब्दावली संग्रह, एवं व्याकरणिक विवरण को डिजिटल माध्यम में सुरक्षित रखकर भाषाई विविधता के संरक्षण में योगदान दिया जा सकता है।

चुनौतियाँ: वास्तविकताओं का कठिन धरातल

संसाधन अकाल (Resource Scarcity) - "डेटा ही राजा है" यह सबसे बड़ी एवं मूलभूत बाधा है। आधुनिक, डेटा-चालित AI/ML मॉडल प्रशिक्षण के लिए विशाल एवं उच्च-गुणवत्ता वाले डेटा पर निर्भर करते हैं।

---

[8] सांख्यिकीय मशीन अनुवाद (SMT) शब्दों और वाक्यांशों के आँकड़ों पर आधारित है, जो उन्हें तोड़कर अनुवाद करता है, जबकि न्यूरल मशीन अनुवाद (NMT) एक बड़े तन्त्रिका नेटवर्क का उपयोग करता है और पूरे वाक्यों को एक साथ अनुवादित करता है, जिससे यह अधिक स्वाभाविक और सन्दर्भित लगता है, जो आज के मशीन अनुवाद में सबसे प्रभावी तरीका है।

[9] भाषा विशिष्ट ध्वनि संग्रह (Speech Corpora) ऑडियो रिकॉर्डिंग और उनके लिखित प्रतिलेखन का एक बड़ा डेटाबेस होता है, जो किसी खास भाषा (जैसे हिंदी, तमिल) की ध्वनियों, उच्चारण और बोली के पैटर्न को इकट्ठा करता है, जिसका उपयोग AI, वाक् पहचान और भाषा विज्ञान के शोध के लिए होता है।

[10] ध्वनि मॉडल (Acoustic Models) मुख्य रूप से स्वचालित वाक् पहचान (ASR) प्रणालियों में उपयोग होते हैं, जो ध्वनि और भाषाई इकाइयों (जैसे अक्षर या ध्वनि) के बीच सांख्यिकीय संबंध स्थापित करते हैं, ताकि यह पता चल सके कि कौन सी आवाज़ किस शब्द या ध्वनि से मेल खाती है, जैसे हिडन मार्कोव मॉडल (HMM) या तंत्रिका नेटवर्क। यह मॉडल भाषण को समझने के लिए ऑडियो और टेक्स्ट के बीच पुल का काम करते हैं, जिससे फ़ोन असिस्टेंट या डिक्टेशन सॉफ्टवेयर काम कर पाते हैं।

[11] ऑप्टिकल कैरेक्टर रिकॉग्निशन (OCR) एक तकनीक है जो किसी इमेज (जैसे स्कैन किए गए दस्तावेज़, फोटो या PDF) में मौजूद टेक्स्ट को पहचानकर उसे मशीन-पठनीय और सम्पादन योग्य डिजिटल टेक्स्ट में बदल देती है, जिससे डेटा को निकालना, खोजना और उपयोग करना आसान हो जाता है; यह स्कैनिंग, पहचान और टेक्स्ट निकालने जैसे चरणों के माध्यम से काम करती है और डेटा एंट्री, डिजिटलीकरण और सुलभता में महत्वपूर्ण भूमिका निभाती है।

[12] टेक्स्ट माइनिंग , टेक्स्ट डेटा माइनिंग (TDM) या टेक्स्ट एनालिटिक्स, टेक्स्ट से उच्च-गुणवत्ता वाली जानकारी प्राप्त करने की प्रक्रिया है। इसमें "कम्प्यूटर द्वारा विभिन्न लिखित संसाधनों से स्वचालित रूप से जानकारी निकालकर नई, पहले से अज्ञात जानकारी की खोज" शामिल है।

एनोटेटेड कोष का अभाव - किसी भाषा के लिए एक उन्नत NLP पाइपलाइन (जैसे **POS Tagger**[13], **Parser**[14], **Named Entity Recognizer**[15]) बनाने के लिए हजारों लाखों वाक्यों के ऐसे संग्रह की आवश्यकता होती है, जिन पर मानव विशेषज्ञों ने व्याकरणिक जानकारी दी हों। उदाहरण के लिए, प्रत्येक शब्द को सञ्ज्ञा, क्रिया आदि के रूप में चिह्नित करना (POS Tagging)। अंग्रेजी के पास **Penn Treebank**[16] जैसे विशाल एनोटेटेड कोष हैं, परन्तु भारतीय भाषाओं के लिए ऐसे संसाधन बहुत सीमित, असङ्गठित एवं अप्रकाशित हैं।

डोमेन-विशिष्ट डेटा का अभाव - चिकित्सा, कानून, तकनीकी जैसे विशिष्ट क्षेत्रों के लिए भारतीय भाषाओं का तैयार डेटा सेट्स का लगभग पूर्ण अभाव है, जिससे डोमेन विशिष्ट अनुप्रयोगों का विकास मुश्किल है।

भाषाई मानकीकरण का अभाव - एक ही भाषा के विभिन्न क्षेत्रीय रूपों, लिपि विविधताओं एवं अनौपचारिक ऑनलाइन लेखन शैलियों के कारण डेटा एकरूप नहीं होता, जिससे मॉडल की सटीकता प्रभावित होती है।

तकनीकी एवं संरचनात्मक बाधाएँ

लिपि सम्बन्धी जटिलताएँ - यूनिकोड मानक के बाद स्थिति बेहतर हुई है, परन्तु मल्टीपल एन्कोडिंग (एक ही अक्षर के लिए विभिन्न कोड पॉइंट्स), फॉन्ट रूपान्तरण, एवं हस्तलिखित पाठ के OCR में अभी भी चुनौतियाँ बनी हुई हैं।

भाषाई जटिलताएँ - भारतीय भाषाओं की समृद्धता (जैसे - ध्वन्यात्मकता, व्याकरणिक विविधता, सांस्कृतिक सन्दर्भ) मशीनों के लिए चुनौती बनती है क्योंकि AI को इन भाषाओं के लिए पर्याप्त डेटा, मानकीकरण की कमी, जटिलता (जैसे संस्कृत या तमिल के मुहावरे), और देवनागरी जैसी लिपियों में लिखने की प्रक्रिया की जटिलता के कारण समझने, प्रोसेस करने और उपयुक्त अनुवाद करने में समस्या आती है।

मुक्त शब्दक्रम - अंग्रेजी में 'राम ने सीता को देखा' का एक ही क्रम हो सकता है। हिन्दी/संस्कृत में 'राम ने सीता को देखा', 'सीता को राम ने देखा', 'देखा राम ने सीता को' आदि सभी व्याकरणिक रूप से सही हैं। इसन दृष्टि से भी भारतीय भाषाओं की वाक्य संरचना का कार्य अत्यन्त जटिल हो जाता है।

शून्य निर्देश - वाक्य में प्रत्यक्ष रूप से उल्लेख किए बिना ही सर्वनामों का प्रयोग। जैसे "[राम] बाजार गया। [स:] पुस्तकं पठति।" मशीन के लिए हर वाक्य में कर्ता/कर्म आदि का पता लगाना कठिन हो जाता है।

पश्चिम-केन्द्रित उपकरणों की सीमा - अधिकांश प्रचलित NLP लाइब्रेरियाँ (जैसे **NLTK, spaCy**[17]**)** एवं प्री-ट्रेण्ड मॉडल (जैसे **BERT**[18]) प्राथमिक रूप से अंग्रेजी एवं यूरोपीय भाषाओं के डेटा पर प्रशिक्षित हैं। भारतीय भाषाओं की विशिष्टताओं के लिए इन्हें शून्य से पुनः प्रशिक्षित करने या नए मॉडल विकसित करने की आवश्यकता है।

---

[13] POS Tagger एक सॉफ्टवेयर या टूल है जो किसी भी लिखे हुए टेक्स्ट (पाठ) को पढ़ता है और हर शब्द को उसकी व्याकरणिक भूमिका के आधार पर एक टैग देता है, जैसे संज्ञा, क्रिया, विशेषण आदि, ताकि कम्प्यूटर भाषा को बेहतर ढंग से समझ सके और प्राकृतिक भाषा प्रसंस्करण (NLP) के विभिन्न कार्यों (जैसे मशीन ट्रांसलेशन, सेंटीमेंट एनालिसिस) में उसका उपयोग कर सके। यह शब्दों को उनके सन्दर्भ और व्याकरण के अनुसार वर्गीकृत करता है।

[14] Parser एक ऐसा सॉफ्टवेयर प्रोग्राम है जो किसी इनपुट डेटा (जैसे टेक्स्ट, कोड, या मैसेज) को लेता है और उसे छोटे, समझने योग्य हिस्सों में तोड़कर एक संरचित प्रारूप में बदलता है, जैसे कि एक ट्री स्ट्रक्चर, ताकि कम्प्यूटर उसे प्रोसेस कर सके और उसके व्याकरणिक नियमों और संरचना को समझ सके।

[15] Named Entity Recognition प्राकृतिक भाषा प्रसंस्करण (NLP) की एक तकनीक है, जिसका उपयोग टेक्स्ट में महत्वपूर्ण जानकारी (इकाइयों) की पहचान करने और उन्हें पूर्वनिर्धारित श्रेणियों में वर्गीकृत करने के लिए किया जाता है।

[16] Penn Treebank एक बहुत बड़ा और महत्वपूर्ण annotated कोष है, जो वाक्यों के वाक्य-विन्यास और अर्थ-संबंधी संरचना को दर्शाता है, जिसमें शब्दों को पार्ट-ऑफ-स्पीच टैग्स और वाक्य-विन्यास वृक्ष (parse trees) के साथ एनोटेट किया गया है, जिससे प्राकृतिक भाषा प्रसंस्करण (NLP) के लिए यह एक आधारभूत संसाधन बन गया है।

[17] NLTK और spaCy दोनों ही लोकप्रिय नेचुरल लैंग्वेज प्रोसेसिंग (NLP) लाइब्रेरीज़ हैं, लेकिन ये अलग-अलग उद्देश्यों को पूरा करती हैं। NLTK मुख्य रूप से एक रिसर्च और शैक्षिक टूलकिट है, जबकि spaCy को परफॉरमेंस और प्रोडक्शन-ग्रेड एप्लिकेशन (औद्योगिक अनुप्रयोगों) के लिए डिज़ाइन किया गया है।

[18] BERT का पूर्ण रूप Bidirectional Encoder Representations from Transformers है।
प्री-ट्रेन्ड (Pre-trained) - इसका मतलब है कि BERT को शुरुआत में बहुत बड़े, बिना लेबल वाले टेक्स्ट डेटासेट (जैसे कि सम्पूर्ण Wikipedia) पर प्रशिक्षित किया गया था। इस व्यापक प्रशिक्षण के दौरान, मॉडल ने भाषा के सामान्य पैटर्न, व्याकरण और सन्दर्भ को समझना सीखा।

मानव संसाधन एवं संस्थागत चुनौतियाँ

अन्तःविषयक अन्तराल - भाषाविज्ञान के विद्वान अक्सर प्रोग्रामिंग एवं एल्गोरिदम से अनभिज्ञ होते हैं, जबकि कम्प्यूटर वैज्ञानिकों को भाषा की सूक्ष्मताओं का गहन ज्ञान नहीं होता। इस संचार की खाई के कारण प्रभावी सहयोग स्थापित होना कठिन है।

शैक्षणिक-औद्योगिक दूरी - विश्वविद्यालयों में होने वाला शोध अक्सर प्रकाशन-केन्द्रित एवं सैद्धान्तिक होता है, जबकि उद्योग को त्वरित, मापनीय एवं उपयोगकर्ता-अनुकूल समाधानों की आवश्यकता होती है। इस दूरी के कारण शोध का व्यावसायिक अनुप्रयोग सीमित रह जाता है।

वित्तपोषण एवं नीतिगत प्राथमिकता - राष्ट्रीय स्तर पर भाषा प्रौद्योगिकी को विज्ञान एवं प्रौद्योगिकी की मुख्यधारा में उचित प्राथमिकता नहीं मिली है। परियोजनाएँ खण्डित एवं अल्पकालिक होती हैं।

भविष्य का मार्ग: एक समन्वित रोडमैप

उपरोक्त चुनौतियों के समाधान हेतु एक बहु-स्तरीय, सहयोगात्मक एवं दीर्घकालिक रणनीति आवश्यक है।

राष्ट्रीय भाषा प्रौद्योगिकी मिशन (National Language Technology Mission - NLTM) की स्थापना - सरकार, उद्योग एवं शिक्षा जगत के साझा प्रयास से एक सार्वजनिक-निजी साझेदारी (PPP) मॉडल पर आधारित एक शीर्ष संस्थान की स्थापना की जानी चाहिए। जिसका उद्देश्य निम्नलिखित हो -

१) भाषाई डेटा अवसंरचना का निर्माण - सभी प्रमुख भारतीय भाषाओं के लिए मानकीकृत, उच्च-गुणवत्ता वाले, खुले एनोटेटेड कोषों का निर्माण एवं रखरखाव।

२) खुले स्रोत के उपकरणों का भण्डार - मूलभूत NLP उपकरणों (Tokenizers, Stemmers, POS Taggers, Parsers) का एक सामान्य, पुनः प्रयोज्य कोडबेस तैयार करना, जिसे शोधकर्ता एवं उद्यमी आसानी से अपनी आवश्यकतानुसार उपयोग कर सकें।

३) मानकीकरण एवं बेंचमार्किंग - लिपि, एनोटेशन योजनाओं, मूल्यांकन मेट्रिक्स आदि के लिए राष्ट्रीय मानक निर्धारित करना।

शिक्षण एवं कौशल विकास में क्रान्ति

अन्तःविषयक पाठ्यक्रम - विश्वविद्यालय स्तर पर 'कम्प्यूटेशनल लिंग्विस्टिक्स' या 'भाषा प्रौद्योगिकी' में स्नातक एवं स्नातकोत्तर डिग्री/डिप्लोमा कार्यक्रम शुरू किए जाएँ। इनमें भाषाविज्ञान, कम्प्यूटर विज्ञान, गणित एवं आँकड़ा विज्ञान का सम्मिलित अध्ययन हो।

प्रशिक्षण कार्यशालाएँ एवं हैकाथॉन्स - छात्रों, शोधकर्ताओं एवं उद्यमियों को खुले डेटासेट्स एवं टूलकिट्स का प्रयोग करने हेतु प्रशिक्षित करना।

उद्योग इण्टर्नशिप - छात्रों को AI/ML कम्पनियों में व्यावहारिक अनुभव प्रदान करना।

अनुसंधान के नवीन क्षेत्रों पर ध्यान केन्द्रित

लघु एवं शून्य-संसाधन भाषाओं के लिए NLP - बहुत कम डिजिटल डेटा वाली भाषाओं के लिए मॉडल विकसित करने की तकनीकों पर शोध। जैसे – क्रॉस-लिंगुअल ट्रांसफर लर्निंग (एक भाषा के ज्ञान का दूसरी भाषा में स्थानान्तरण), अनसुपरवाइज्ड लर्निंग।

मल्टीमॉडल एवं मल्टीटास्किंग मॉडल्स - पाठ, आवाज, चित्र (जैसे हस्तलेख) को एक साथ प्रसंस्कृत करने वाले मॉडलों का विकास।

नैतिक AI एवं भाषाई पक्षपात - AI मॉडलों में निहित लिङ्गगत, जातिगत या क्षेत्रीय पूर्वाग्रहों की पहचान एवं निवारण पर कार्य करना।

अन्तर्राष्ट्रीय सहयोग को बढ़ावा

यूरोपीय संघ की CLARIN (भाषाई संसाधनों के लिए शोध अवसंरचना) जैसी पहलों से सीखकर, भारत एशिया या ब्रिक्स देशों के साथ मिलकर बहुभाषिक एवं बहु-लिपि प्रौद्योगिकी के लिए साझा मंच विकसित कर सकता है।

निष्कर्ष

द्विदिशात्मक - BERT एक वाक्य में शब्दों को दोनों दिशाओं (बाएं से दाएं और दाएं से बाएं) से एक साथ पढ़कर सन्दर्भ को समझता है। यह सुविधा इसे अन्य पुराने मॉडलों से बेहतर बनाती है, जो केवल एक ही दिशा से पढ़ते थे।

ट्रांसफॉर्मर - BERT ट्रांसफॉर्मर नामक एक आधुनिक न्यूरल नेटवर्क आर्किटेक्चर का उपयोग करता है, जो शब्दों के बीच के जटिल सम्बन्धों को प्रभावी ढंग से पकड़ने में मदद करता है।

कम्प्यूटर युग भारतीय भाषाविज्ञान के लिए एक पुनर्जागरण का काल है। एक ओर, इसकी प्राचीन, एल्गोरिदमिक परम्पराएँ वैश्विक AI शोध को नई दिशा दे सकती हैं। दूसरी ओर, अपनी जीवन्त बहुभाषिकता के कारण, यह प्रौद्योगिकी नवाचार का एक विशाल प्रयोगशाला क्षेत्र है। हालांकि, इस मार्ग में संसाधनों का अकाल, तकनीकी जटिलताएँ एवं संस्थागत कमजोरियाँ जैसी गम्भीर बाधाएँ हैं। इन बाधाओं का समाधान तभी सम्भव है जब भाषाविद्, कम्प्यूटर वैज्ञानिक, नीति निर्माता एवं उद्योग जगत एक मंच पर आएँ। राष्ट्रीय भाषा प्रौद्योगिकी मिशन जैसी एक सुदृढ़ संस्थागत संरचना, अन्तःविषयक शिक्षा के माध्यम से युवा मस्तिष्कों का निर्माण, एवं खुले स्रोत के सहयोगात्मक विकास की संस्कृति ये तीन स्तम्भ भारतीय भाषा विज्ञान को डिजिटल युग में नेतृत्वकारी भूमिका के लिए तैयार कर सकते हैं। तब यह केवल अतीत के गौरव का स्मरण नहीं, बल्कि भविष्य की डिजिटल दुनिया का एक सक्रिय निर्माता बन सकेगा। यह केवल भाषाई प्रगति नहीं, बल्कि भारत के सामाजिक तथा आर्थिक विकास एवं सांस्कृतिक पहचान को इस डिजिटल काल में सुरक्षित रखने की दिशा में एक निर्णायक कदम होगा।

सन्दर्भ ग्रन्थसूची

1. शर्मा, आर्येन्द्र, (2008), संस्कृत साहित्य और कम्प्यूटर, नई दिल्ली: इन्दिरा गांधी राष्ट्रीय कला केन्द्र।
2. झा, गिरीश नाथ, (2015), द फिलॉसफी ऑफ़ संस्कृत ग्रामर एण्ड इट्स रिलेवेंस टू कम्प्यूटेशनल लिंग्विस्टिक्स, वाराणसी. संस्कृत विश्वविद्यालय प्रकाशन।
3. कुलकर्णी, अम्बा, (2020), संस्कृत कम्प्यूटेशनल लिंग्विस्टिक्स. अ प्राइमर, नई दिल्ली. डी.के. प्रिंटवर्ल्ड।
4. भारत सरकार, (2022). नेचुरल लैंग्वेज टेक्नोलॉजी मिशन (NLTM) विज़न डॉक्यूमेंट, इलेक्ट्रॉनिक्स और सूचना प्रौद्योगिकी मन्त्रालय (MeitY)।
5. कपूर, नायर एवं अन्य, (2023). "IndicBERT: A Pre-trained Language Model for Indian Languages" ACL Anthology.
6. भट्टाचार्य, पुष्पक, (2019), डिजिटल ह्यूमैनिटीज एण्ड इंडोलॉजी: न्यू डायरेक्शन्स, चेन्नई: यूनिवर्सिटी ऑफ़ मद्रास प्रेस।
7. चॉम्स्की, नोम, (1965). आस्पेक्ट्स ऑफ द थ्योरी ऑफ सिंटेक्स, एमआईटी प्रेस।
8. गोयल, पवन एवं अन्य, (2016). "Computational Structure of the Ashtadhyayi". Sanskrit Computational Linguistics Symposium.
9. हुइत, पी, (2009). संस्कृत सिंटेक्स एंड कम्प्यूटर प्रोसेसिंग. पेरिस: स्प्रिंगर।
10. राष्ट्रीय शिक्षा नीति 2020. भाषा और संस्कृति संरक्षण अनुभाग. भारत सरकार।
11. सिंह, अनिल कुमार. (2024). आर्टिफिशियल इंटेलिजेंस और भारतीय भाषाएँ. प्रयागराज: हिन्दी साहित्य सम्मेलन।
12. पाणिनि. अष्टाध्यायी. टीकाकार: वामन एवं जयादित्य, काशिका वृत्ति।