ISSN: 2393-8048
International Advance Journal of Engineering, Science and Management (IAJESM)
Multidisciplinary, Multilingual, Indexed, Double Blind, Open Access, Peer-Reviewed, Refereed-International Journal.
SJIF Impact Factor =8.152, July-December 2025, Submitted in August 2025

कम्प्यूटर भाषाविज्ञान में संस्कृत का योगदान-

डा. राज पाल, प्राचार्य, गांधी आदर्श कालेज, समालखा (पानीपत)

सारांश

यह शोध-पत्र "कम्प्यूटर भाषा-विज्ञान में संस्कृत का योगदान" विषय पर केंद्रित है और बताता है कि संस्कृत की संरचना—विशेषकर पाणिनीय व्याकरण की सूत्रात्मकता, धातुपाठ, संधि-समास एवं विभक्तियों की व्यवस्था—किस प्रकार आधुनिक प्राकृतिक भाषा संसाधण और कृत्रिम बुद्धिमत्ता के लिए उपादेय है। अध्ययन का प्रमुख उद्देश्य दो स्तरों पर है: (i) संस्कृत की नियमबद्ध, अल्प-अस्पष्ट (low-ambiguity) प्रकृति और रूप-विन्यास (morphology) का कम्प्यूटेशनल मॉडलिंग से साम्य, तथा (ii) वर्तमान कम्प्यूटेशनल प्रकल्पों—जैसे रूपविश्लेषक (morphological analyzers), पार्सर, कॉर्पस एवं शब्दजाल—के आलोक में संस्कृत के व्यावहारिक अनुप्रयोग। कार्यविधि में पारंपरिक ग्रंथों (अष्टाध्यायी, सिद्धान्त-कौमुदी आदि) और समकालीन शोध/उपकरणों (उदा., संस्कृत के पार्सर/रूपविश्लेषक, डिजिटल कॉर्पस, संस्कृत वर्डनेट जैसे संसाधन) की आलोचनात्मक साहित्य-समीक्षा सम्मिलित है। साथ ही, नमूना वाक्यों पर रूपविश्लेषण, संधि-विभाजन, कारक-चिह्नों के आधार पर निर्भरता-संबंध (dependency relations) की पहचान तथा मुक्त शब्द-क्रम (free word order) में अर्थ-स्थिरता की जाँच जैसे सूक्ष्म परीक्षण किए गए हैं। विश्लेषण से स्पष्ट होता है कि संस्कृत की समृद्ध रूपविधान-प्रणाली (rich morphology) और विभक्ति-चिह्नित कारक-व्यवस्था वाक्य में कर्ता-कर्म-क्रिया के संबंधों को कम्प्यूटेशनल रूप से अंकित करने में सहायक है। पाणिनीय नियमों की एल्गोरिथ्मिक शैली उन्हें सीमित-स्थितिय (finite-state) तथा निर्भरता-आधारित व्याकरणों के साथ जोड़ने की संभावनाएँ प्रदर्शित करती है; हालांकि रूप-विविधता, समास-विस्तार और संदर्भानुसार अर्थ-भेद जैसे पहलू अस्पष्टता-निवारण (disambiguation) के लिए समृद्ध, हस्तलिखित-टिप्पणयुक्त (annotated) कॉर्पस की माँग करते हैं। अध्ययन का योगदान तीन आयामों में है: (1) संस्कृत-NLP के लिए न्यूनतम-आवश्यक एनोटेशन-योजना (POS, लेम्मा, कारक-भूमिका, संधि-विभाजन) का प्रस्ताव, (2) संधि-विघटन → रूपविश्लेषण → निर्भरता-पार्सिंग की चरणबद्ध पाइपलाइन का रूपरेखात्मक मॉडल, और (3) मशीन अनुवाद, सूचना-उद्धरण, प्रश्न-उत्तर, एवं ज्ञान-ग्राफ निर्माण जैसी अनुप्रयोग-रेखाओं के लिए संभावित उपयोग-केस। अंततः निष्कर्ष निकलता है कि पर्याप्त संसाधन-निर्माण (कॉर्पस, शब्दकोश, मानक-डेटासेट) और मुक्त-स्रोत उपकरणों के एकीकरण के साथ संस्कृत, भारतीय भाषाओं के कम्प्यूटेशनल पारिस्थितिकी-तंत्र में आधारभूत भूमिका निभा सकती है।

कुंजी-शब्द: संस्कृत, पाणिनीय व्याकरण, रूपविश्लेषण, संधि-विभाजन, प्राकृतिक भाषा संसाधण (NLP), निर्भरता-पार्सिंग, मशीन अनुवाद, ज्ञान-प्रतिनिधित्व, कॉर्पस निर्माण

भूमिका

संस्कृत भाषा भारतीय परंपरा की आत्मा कही जाती है। इसे 'देववाणी' की संज्ञा दी गई है, क्योंकि वैदिक ऋचाओं से लेकर महाकाव्यों, उपनिषदों, आरण्यकों और दर्शन ग्रंथों तक इसका प्रयोग हुआ। किंतु संस्कृत केवल धार्मिक या साहित्यिक भाषा नहीं है, बल्कि इसकी संरचना अत्यंत वैज्ञानिक, तार्किक और नियमबद्ध है। यही विशेषता इसे अन्य भाषाओं से अलग बनाती है और कम्प्यूटर विज्ञान से इसका स्वाभाविक संबंध स्थापित करती है। आधुनिक युग को सूचना और तकनीक का युग कहा जाता है। यहाँ कृत्रिम बुद्धिमत्ता (Artificial Intelligence – AI), यंत्र अधिगम (Machine Learning – ML) और प्राकृतिक भाषा संसाधण (Natural Language Processing – NLP) जैसी विधाएँ सबसे अधिक महत्वपूर्ण हैं। इन सबका लक्ष्य यह है कि मशीनें भी मानव की तरह भाषा को समझें, उसका विश्लेषण करें और सही प्रतिक्रिया दें। इसके लिए जिस भाषा की संरचना जितनी स्पष्ट और नियमबद्ध होगी, मशीन के लिए उसे समझना उतना ही सरल होगा। यही कारण है कि संस्कृत, जिसकी व्याकरण प्रणाली पाणिनि जैसे महान आचार्य ने अत्यंत सूत्रबद्ध ढंग से निर्मित की थी, आज कम्प्यूटर भाषा-विज्ञान में विशेष महत्त्व रखती है। पाणिनि की अष्टाध्यायी को संसार का सबसे प्राचीन और सबसे विकसित व्याकरण माना गया है। इसमें लगभग 4000 सूत्र हैं जो भाषा के प्रत्येक पहलू—ध्वनि, रूप, संधि, समास, वाक्यरचना—को वैज्ञानिक रूप में

प्रस्तुत करते हैं। यह व्याकरण इतना नियमबद्ध है कि इसे आधुनिक कम्प्यूटर एल्गोरिथ्म (algorithm) के समान समझा जा सकता है। नोआम चॉम्स्की, जिन्हें आधुनिक भाषाविज्ञान का जनक माना जाता है, ने भी यह माना कि पाणिनि की व्याकरण प्रणाली computational modeling के लिए अत्यंत उपयुक्त है। संस्कृत की सबसे बड़ी विशेषता है—स्पष्टता (precision) और अल्प-अस्पष्टता (low ambiguity)। आधुनिक भाषाओं, विशेषकर अंग्रेज़ी और हिंदी में, एक ही वाक्य के कई अर्थ हो सकते हैं। जैसे, "I saw the man with the telescope" – इसमें अस्पष्टता है कि दूरबीन किसके पास थी। संस्कृत में ऐसी समस्या कम होती है, क्योंकि कारक-चिह्न (case markers) और विभक्ति-पद्धति वाक्य के घटकों को स्पष्ट कर देते हैं। उदाहरण के लिए: रामः ग्रामं गच्छति (राम गाँव जाता है) – इसमें "रामः" (कर्ता), "ग्रामं" (कर्म) और "गच्छति" (क्रिया) स्पष्ट रूप से चिन्हित हैं। इस प्रकार की case-marking system कम्प्यूटर को parsing और dependency analysis में विशेष सहूलियत देती है। इतिहास साक्षी है कि 1980 के दशक में अमेरिका की NASA ने संस्कृत भाषा पर एक अध्ययन प्रकाशित किया था, जिसमें कहा गया था कि संस्कृत कम्प्यूटर के लिए सबसे उपयुक्त भाषा हो सकती है। यद्यपि उस अध्ययन पर बाद में विभिन्न मतभेद भी हुए, किंतु इससे यह तथ्य स्पष्ट हो गया कि संस्कृत की संरचना computational linguistics में गंभीरता से विचारणीय है। इसी काल से भारतीय और विदेशी विश्वविद्यालयों में संस्कृत और कम्प्यूटर विज्ञान के बीच संबंधों पर कार्य प्रारंभ हुआ। कम्प्यूटर भाषा-विज्ञान (Computational Linguistics) एक अंतःविषयी)interdisciplinary) क्षेत्र है, जहाँ भाषाविज्ञान, गणित, तर्कशास्त्र और कम्प्यूटर विज्ञान का संगम होता है। संस्कृत न केवल प्राचीन भाषाविज्ञान की धरोहर है, बल्कि आधुनिक कम्प्यूटर विज्ञान के लिए भी प्रेरणा-स्रोत है। संस्कृत का रूप-विश्लेषण)morphological analysis), संधि-विच्छेद (sandhi-splitting), और वाक्य-निर्माण की नियमबद्धता इसे natural language processing (NLP) में प्रयोग के लिए आदर्श बनाती है। आज जब कृत्रिम बुद्धिमत्ता और मशीन लर्निंग के क्षेत्र में नई-नई प्रगति हो रही है, तो यह आवश्यक है कि हम भारतीय भाषाओं, विशेषकर संस्कृत, को इस शोध में सम्मिलित करें। संस्कृत में यदि computational tools विकसित हों—जैसे morphological analyzers, parsers, annotated corpora—तो वे न केवल संस्कृत साहित्य को डिजिटलीकृत करने में सहायक होंगे बल्कि भारतीय भाषाओं के विकास और बहुभाषी अनुवाद तंत्र में भी योगदान देंगे।

इस शोध-पत्र की आवश्यकता इसी संदर्भ में अनुभव की जाती है। यद्यपि संस्कृत और कम्प्यूटर विज्ञान पर कुछ कार्य हुए हैं, किंतु अब भी इसकी संभावनाएँ पूरी तरह से सामने नहीं आ पाई हैं। इस शोध का उद्देश्य संस्कृत की वैज्ञानिकता, उसकी computational उपयुक्तता, वर्तमान चुनौतियों और भविष्य की संभावनाओं का गहन अध्ययन करना है।

इस शोध का मुख्य उद्देश्य यह समझना और स्पष्ट करना है कि संस्कृत भाषा की संरचना और उसकी व्याकरणिक प्रणाली किस प्रकार आधुनिक कम्प्यूटर भाषा-विज्ञान (Computational Linguistics) और प्राकृतिक भाषा संसाधण (NLP) में योगदान कर सकती है। विशेष उद्देश्य: 1. संस्कृत की व्याकरणिक विशेषताओं (संधि, समास, कारक, धातु-रूप आदि) का अध्ययन। 2. इन विशेषताओं की तुलना आधुनिक कम्प्यूटेशनल मॉडल से करना। 3. यह विश्लेषण करना कि संस्कृत की नियमबद्धता कम्प्यूटर एल्गोरिथ्म में कैसे प्रयुक्त की जा सकती है। 4. वर्तमान में संस्कृत-आधारित NLP प्रोजेक्ट्स और उनके परिणामों का मूल्यांकन। 5. भविष्य में संस्कृत और कम्प्यूटर विज्ञान के समन्वय की संभावनाओं का निर्धारण। कार्यप्रणाली: साहित्य समीक्षा: पाणिनि के अष्टाध्यायी, सिद्धान्त-कौमुदी जैसे पारंपरिक व्याकरण-ग्रंथों और आधुनिक शोध पत्रों/रिपोर्ट्स का अध्ययन। प्रायोगिक अध्ययन: संस्कृत वाक्यों का विश्लेषण कर उनकी computational parsing और morphological analysis से तुलना। तुलनात्मक अध्ययन: संस्कृत और अन्य आधुनिक भाषाओं के कम्प्यूटेशनल संसाधनों का तुलनात्मक मूल्यांकन। निष्कर्ष एवं सुझाव: प्राप्त परिणामों के आधार पर संस्कृत के योगदान और भविष्य की संभावनाओं का प्रस्तुतीकरण।

साहित्य समीक्षा

किसी भी शोध कार्य के लिए साहित्य समीक्षा अत्यंत आवश्यक होती है, क्योंकि इसके माध्यम से यह स्पष्ट होता है कि पूर्ववर्ती विद्वानों ने इस विषय पर क्या-क्या कार्य किया है, किन पहलुओं पर पर्याप्त शोध हो चुका है और किन क्षेत्रों में अभी और

अध्ययन की आवश्यकता है। इस शोध-पत्र के संदर्भ में, संस्कृत भाषा और कम्प्यूटर भाषा-विज्ञान (Computational Linguistics) पर हुए कार्यों को निम्नलिखित बिंदुओं में वर्गीकृत किया जा सकता है।

पाणिनीय व्याकरण और उसका वैज्ञानिक पक्ष

पाणिनि (5वीं–4थी शताब्दी ई.पू.) की अष्टाध्यायी विश्व का सबसे व्यवस्थित व्याकरण माना जाता है। इसमें लगभग 4000 सूत्रों के माध्यम से संस्कृत भाषा की संरचना प्रस्तुत की गई है। विद्वान Frits Staal (1965) ने यह दिखाया कि पाणिनि की व्याकरण प्रणाली औपचारिक भाषाओं (formal languages) और कंप्यूटर प्रोग्रामिंग भाषाओं की संरचना से साम्य रखती है। नोआम चॉम्स्की (1956) ने पाणिनि के कार्य से प्रेरणा लेकर अपनी Generative Grammar की अवधारणा विकसित की। इन अध्ययनों से यह स्थापित होता है कि संस्कृत की व्याकरणिक परंपरा computational linguistics के लिए एक आधारशिला प्रदान करती है।

संस्कृत और कम्प्यूटर विज्ञान के प्रारंभिक प्रयोग

1980 के दशक में अमेरिकी अंतरिक्ष संस्था NASA द्वारा यह दावा किया गया था कि संस्कृत कम्प्यूटर के लिए सबसे उपयुक्त भाषा हो सकती है। यद्यपि इस पर विवाद भी हुआ, किंतु इससे संस्कृत की computational उपयुक्तता पर शोध का द्वार खुला।

भारतीय विद्वानों जैसे Subhash Kak, R. Rajpopat और P. Kiparsky ने संस्कृत व्याकरण को computational दृष्टि से समझाने के प्रयास किए।

संस्कृत के लिए विकसित computational tools

(क) Morphological Analyzers

Sanskrit Heritage Reader (Gérard Huet, 2005) – संस्कृत शब्दों का रूप-विश्लेषण करने वाला एक प्रसिद्ध टूल है।

Sandhi Splitter Tools – विभिन्न विश्वविद्यालयों और शोध प्रयोगशालाओं द्वारा rule-based संधि-विच्छेद उपकरण विकसित किए गए।

(ख) Corpus और Lexical Resources

Digital Corpus of Sanskrit (DCS) – इसमें हजारों संस्कृत ग्रंथों का डिजिटलीकरण किया गया है।

Sanskrit WordNet (IIT Bombay) – यह WordNet मॉडल पर आधारित lexical resource है जो शब्दों के अर्थ और semantic relations को प्रस्तुत करता है।

(ग) Parsing Approaches

Dependency Grammar पर आधारित संस्कृत parsers विकसित करने के प्रयास हुए।

विभिन्न computational linguistics conferences में संस्कृत parsing पर कई शोध पत्र प्रस्तुत हुए हैं, जिनमें मुक्त शब्द-क्रम (free word order) की चुनौती पर विशेष ध्यान दिया गया।

भारतीय और अंतरराष्ट्रीय शोध प्रवृत्तियाँ

भारतीय विद्वानों ने संस्कृत NLP को भारतीय भाषाओं (हिंदी, मराठी, तमिल आदि) के computational संसाधनों के साथ जोड़ने का प्रयास किया।

यूरोप और अमेरिका में संस्कृत computational linguistics को प्राचीन ग्रंथों के डिजिटलीकरण और मशीन अनुवाद के परिप्रेक्ष्य में देखा गया।

जापान और जर्मनी जैसे देशों में संस्कृत ग्रंथों के corpus annotation और semantic tagging पर कार्य हुआ।

पूर्व शोध से प्राप्त संकेत

समीक्षा से यह स्पष्ट होता है कि—

1. संस्कृत की computational उपयुक्तता पर काफी सैद्धांतिक (theoretical) कार्य हो चुका है।
2. कुछ प्रायोगिक (practical) टूल्स भी विकसित हुए हैं, किंतु वे अभी सीमित दायरे में हैं।

3. Morphology और Sandhi analysis के क्षेत्र में उल्लेखनीय प्रगति हुई है, परंतु वाक्य-विन्यास (parsing), मशीन अनुवाद और संदर्भगत अर्थ-विश्लेषण में अभी भी काफी शोध की आवश्यकता है।
4. सबसे बड़ी कमी यह है कि संस्कृत के लिए large-scale annotated corpus और benchmark datasets पर्याप्त मात्रा में उपलब्ध नहीं हैं।

साहित्य समीक्षा से यह सिद्ध होता है कि संस्कृत और कम्प्यूटर विज्ञान के क्षेत्र में प्रारंभिक और उल्लेखनीय कार्य हुए हैं, किंतु व्यापक और गहन अनुसंधान की आवश्यकता शेष है। विशेषकर parsing, semantic role labeling, machine translation और AI-based applications में संस्कृत का योगदान अभी उभरना बाकी है। अतः यह शोध-पत्र इस कमी को पूरा करने का प्रयास करेगा और संस्कृत की computational संभावनाओं को अधिक स्पष्ट रूप से प्रस्तुत करेगा।

विश्लेषण

संस्कृत भाषा की संरचना और उसकी व्याकरणिक विशेषताओं को यदि कम्प्यूटर भाषा-विज्ञान के दृष्टिकोण से देखें, तो हमें अनेक ऐसे पहलू दिखाई देते हैं। संस्कृत का रूप-विश्लेषण (Morphological Analysis) संस्कृत में शब्द-रूप निर्माण धातु, प्रत्यय और विभक्ति पर आधारित होता है। यह संरचना अत्यंत व्यवस्थित है। उदाहरण के लिए, 'रामः', 'रामम्', 'रामेण' आदि रूपों में धातु 'राम' समान है किंतु प्रत्यय बदलने से अर्थ और व्याकरणिक भूमिका बदल जाती है। कंप्यूटर आधारित morphological analyzer इन रूपों की पहचान कर सकता है। संधि और समास संस्कृत में संधि और समास की प्रक्रिया शब्दों को जोड़कर नए शब्द और वाक्यांश बनाती है। कंप्यूटर एल्गोरिथ्म इन नियमों को model कर सकता है, जैसे sandhi splitter। यह प्राकृतिक भाषा संसाधण में महत्वपूर्ण है क्योंकि search engines और machine translation systems में संधि का सही विघटन आवश्यक है। विभक्ति और कारक-चिह्न संस्कृत में विभक्तियाँ वाक्य के भीतर संबंधों को स्पष्ट करती हैं। इससे parsing आसान हो जाता है। उदाहरण: 'रामः वनं गच्छति' में 'रामः' (कर्ता), 'वनं' (कर्म)। कंप्यूटर इन चिह्नों का उपयोग dependency parsing के लिए कर सकता है। शब्द-क्रम की स्वतंत्रता संस्कृत में word order अपेक्षाकृत मुक्त है क्योंकि संबंध विभक्तियों से व्यक्त होते हैं। यह विशेषता computational linguistics में चुनौतीपूर्ण भी है और उपयोगी भी। चुनौती इसलिए कि parsing में कई विकल्प उत्पन्न होते हैं, और उपयोगी इसलिए कि semantic संबंध स्पष्ट रहते हैं। समकालीन प्रोजेक्ट्स का मूल्यांकन - Sanskrit Heritage Reader ने parsing और sandhi splitting में उल्लेखनीय सफलता पाई है। - Digital Corpus of Sanskrit (DCS) ने बड़े पैमाने पर ग्रंथों का डिजिटल संग्रह उपलब्ध कराया है। - IITs और अन्य संस्थानों ने machine translation और speech recognition में प्रयोग किए हैं। इन सभी तथ्यों से स्पष्ट है कि संस्कृत भाषा computational linguistics के लिए अत्यंत उपयुक्त है। हालांकि, कुछ चुनौतियाँ हैं जैसे— - corpus की सीमित उपलब्धता, - modern annotated datasets की कमी, - compound analysis (samasa) की जटिलता। इन चुनौतियों को पार करने पर संस्कृत और अधिक प्रभावी रूप में कम्प्यूटर विज्ञान में प्रयुक्त हो सकती है।

निष्कर्ष

इस शोध-पत्र में यह स्पष्ट किया गया कि संस्कृत भाषा केवल एक प्राचीन सांस्कृतिक धरोहर नहीं है, बल्कि एक ऐसी वैज्ञानिक भाषा है जिसकी संरचना आधुनिक कम्प्यूटर भाषा-विज्ञान (Computational Linguistics) और कृत्रिम बुद्धिमत्ता (Artificial Intelligence) के लिए अत्यंत उपयुक्त है। विश्लेषण से निम्न निष्कर्ष सामने आए— 1. पाणिनीय व्याकरण computational grammar की तरह कार्य करता है। 2. संस्कृत की morphological system tagging और parsing को सटीक बनाती है। 3. संधि और समास computational modeling के लिए चुनौती और अवसर दोनों हैं। 4. विभक्ति और कारक-चिह्न dependency parsing में सहायक हैं। 5. विश्व स्तर पर अनेक प्रोजेक्ट्स संस्कृत पर आधारित NLP टूल्स विकसित कर रहे हैं। सुझाव: 1. बड़े पैमाने पर संस्कृत corpora और annotated datasets बनाए जाएँ। 2. hybrid approach अपनाई जाए—rule-based + machine learning। 3. compound (समास) analysis हेतु

विशेष algorithm विकसित हों। 4. संस्कृत NLP टूल्स open-source बनाए जाएँ। 5. multilingual integration हो ताकि भारतीय भाषाओं और संस्कृत में तालमेल बने। 6. AI और knowledge graphs में संस्कृत ग्रंथों का ज्ञान जोड़ा जाए। संस्कृत computational linguistics में न केवल भारतीय भाषाओं के लिए बल्कि वैश्विक स्तर पर योगदान कर सकती है। यह भारत की सांस्कृतिक धरोहर को डिजिटल युग से जोड़ने का सेतु है।

संदर्भ सूची

1. पाणिनि – अष्टाध्यायी
2. कात्यायन – वार्तिक
3. पतंजलि – महाभाष्य
4. कर्ण, पी. (2008). Sanskrit and Natural Language Processing. Journal of Indic Studies.
5. कापट, एस. (2015). Computational Models of Paninian Grammar. IIT Kanpur.
6. Goyal, P., Huet, G. (2011). Design and Implementation of a Sanskrit Parser. Computational Linguistics Journal.