

## नैसर्गिक भाषा प्रक्रिया आणि भाषिक विश्लेषण: एक सिंहावलोकन

प्रा. संतोष मोहन जाधव

मराठी विभाग प्रमुख

श्री नरेंद्र तिडके कला व वाणिज्य महाविद्यालय,  
रामटेक

ई-मेल : jadhaosm78@gmail.com

भ्रमणध्वनी क्रमांक : ९४०५५८८०१२

Crossref DOI - <https://doi.org/10.63665/rh.v7i2.124>

### सारांश :

नैसर्गिक भाषा प्रक्रिया (एनएलपी) हे संगणक विज्ञान आणि कृत्रिम बुद्धिमत्ता (एआय) चे एक उपक्षेत्र आहे जे संगणकांना मानवी भाषेत समजून घेण्यास आणि संवाद साधण्यास सक्षम करण्यासाठी मशीन लर्निंगचा वापर करते. मोठ्या भाषा मॉडेल्स (एलएलएम) च्या संप्रेषण क्षमतांपासून ते प्रतिमा निर्मिती मॉडेल्सच्या कार्ये समजून घेण्याच्या क्षमतेपर्यंत, कृत्रिम बुद्धिमत्ता (एआयच्या) युगाला शक्य करण्यात एनएलपी संशोधनाने महत्त्वाची भूमिका बजावली आहे. एनएलपी आधीच अनेक लोकांच्या दैनंदिन जीवनाचा एक भाग आहे, जो स्पोकन कमांडसह ग्राहक सेवेसाठी सर्च इंजिन आणि चॅटबॉट्सना शक्ती देतो. एनएलपी व्यावसायिक दृष्टीने देखील वाढती भूमिका बजावत आहे जे व्यवसाय कार्ये सुव्यवस्थित आणि स्वयंचलित करतात, कर्मचारी उत्पादकता वाढवतात आणि व्यवसाय प्रक्रिया सुलभ करतात. एनएलपी वापरकर्त्यांच्या प्रश्नांमागील हेतू समजून घेण्यास सिस्टमला मदत करून शोध सुधारते, ज्यामुळे अधिक अचूक आणि संबंधित परिणाम मिळतात. याच अनुषंगाने प्रस्तुत शोधलेखात नैसर्गिक भाषा प्रक्रिया आणि भाषिक विश्लेषण यावर चिंतन करण्यात आलेले आहे.

**बिजशब्द :** नैसर्गिक भाषा प्रक्रिया (एनएलपी), भाषिक विश्लेषण, संगणक, कृत्रिम बुद्धिमत्ता,

### प्रस्तावना :

भाषा ही मानवी संस्कृतीचा पाया असून संवादाचे सर्वात प्रभावी साधन आहे. डिजिटल युगामध्ये भाषेचे महत्त्व केवळ दैनंदिन संभाषणापुरते मर्यादित राहिलेले नाही, तर ती माहिती तंत्रज्ञान, उद्योग, प्रशासन, शिक्षण आणि सामाजिक माध्यमे या सर्व क्षेत्रांत मध्यवर्ती ठरली आहे. संगणकांना मानवी भाषेचा अर्थ समजून घेता यावा आणि ते योग्य प्रतिसाद देऊ शकावेत, यासाठी नैसर्गिक भाषा प्रक्रिया आणि कृत्रिम बुद्धिमत्ता या दोन शाखांचा समन्वय अत्यंत महत्त्वाचा ठरतो. Jurafsky आणि Martin यांनी Speech and Language Processing (2023,) या पुस्तकात स्पष्ट केले आहे की नैसर्गिक भाषा प्रक्रिया (NLP) म्हणजे संगणक विज्ञान आणि भाषाशास्त्र यांचा संगम आहे, इंग्रजी, चिनी, जपानी अशा भाषांमध्ये NLP विषयक विपुल संशोधन उपलब्ध



आहे; मात्र भारतीय भाषांमध्ये हे कार्य मर्यादित आहे. हिंदीत काही साधने तयार झाली असली तरी मराठीमध्ये या संदर्भातील प्रगती अद्याप प्राथमिक टप्प्यावर आहे. सध्याच्या डिजिटल युगात संगणकीय प्रणाली केवळ आकडेमोडी, डेटा प्रक्रिया किंवा यांत्रिक कार्यांपुरती मर्यादित राहिलेली नाही. आज संगणक मानवी संवाद, भावना, भाषा आणि बोली यांचाही अर्थ लावू शकतात. यामागे एक महत्त्वाची तांत्रिक संकल्पना कार्यरत आहे ती म्हणजे नैसर्गिक भाषा प्रक्रिया (Natural Language Processing – NLP). ही संकल्पना संगणक विज्ञान आणि कृत्रिम बुद्धिमत्ता (AI) या शाखांशी निगडित असून तिचा मुख्य उद्देश संगणकाला मानवी भाषा समजून घेण्यास, तिचे विश्लेषण करण्यास आणि योग्य प्रतिसाद देण्यास सक्षम बनवणे हा आहे. नैसर्गिक भाषा प्रक्रिया ही संगणकाला मानवाच्या नैसर्गिक भाषेतून संवाद साधण्याची क्षमता प्रदान करते. NLP चा वापर करून संगणक प्रणालींना भाषा समजणे (Language Understanding), भाषा निर्माण करणे (Language Generation), आवाज ओळखणे (Speech Recognition), मजकुराचा आवाजात रूपांतरण (Text-to-Speech), भाषांतर (Machine Translation) तसेच भावना विश्लेषण (Sentiment Analysis) यांसारखी कार्ये करता येतात. Bird, Klein आणि Loper या अभ्यासकांनी आपल्या ग्रंथांमध्ये स्पष्ट केले आहे की, NLP चा मूलभूत हेतू म्हणजे मानवी भाषेला मशीनसाठी प्रक्रियायोग्य स्वरूपात आणणे होय.

### नैसर्गिक भाषा संस्करणची प्रमुख वैशिष्ट्ये :

- **नैसर्गिक भाषा समजून घेणे** : नैसर्गिक भाषा संस्करण मानवी भाषा दररोजच्या संभाषणात बोलल्या किंवा लिहिल्या जाणाऱ्या भाषेला समजून घेण्यासाठी डिझाइन केलेले असतात. यामध्ये समानार्थी शब्द, बोलचाल, अपभाषा आणि विविध भाषा संरचना समजून घेणे समाविष्ट आहे.
- **संदर्भ जागरूकता** : नैसर्गिक भाषा संस्करण संबंधित प्रतिसाद देण्यासाठी वापरकर्त्यांच्या इनपुटच्या संदर्भाचा विचार करतात. ते त्यांच्या प्रतिसादांची माहिती देण्यासाठी मागील परस्परसंवाद, वापरकर्ता प्राधान्ये आणि बाह्य डेटा वापरतात.
- **परस्परसंवादी संवाद** : नैसर्गिक भाषा संस्करण परस्परसंवादी संभाषणांमध्ये गुंततात, ज्यामुळे वापरकर्त्यांना पुढील प्रश्न विचारता येतात, प्रतिसाद स्पष्ट करता येतात किंवा संभाषणाचा विषय बदलता येतो.
- **स्पष्टीकरण** : जर वापरकर्ता इनपुट अस्पष्ट असेल किंवा अनेक प्रकारे अर्थ लावता येईल, तर एक चांगला नैसर्गिक भाषा संस्करण वापरकर्त्यांचा हेतू समजून घेण्यासाठी स्पष्टीकरणात्मक प्रश्न विचारेल.
- **अर्थपूर्ण प्रक्रिया** : नैसर्गिक भाषा संस्करण केवळ भाषेच्या वाक्यरचनावर प्रक्रिया करत नाहीत तर ते शब्द आणि वाक्यांशांमागील अर्थाचे विश्लेषण देखील करतात.
- **आवाज-आधारित** : आवाज-आधारित नैसर्गिक भाषेच्या संस्करणच्या बाबतीत, ते बोलल्या जाणाऱ्या भाषेचे लिप्यंतरण करण्यासाठी स्पीच रेकग्निशन आणि प्रतिसादांना तोंडी करण्यासाठी टेक्स्ट-टू-स्पीच तंत्रज्ञानाचा वापर करतात. हे वैशिष्ट्य अधिक चांगल्या प्रकारे समजून घेण्यासाठी, विविध मोफत टेक्स्ट-टू-स्पीच टूलसची चाचणी घेणे फायदेशीर आहे. ते तंत्रज्ञानाचा प्रत्यक्ष अनुभव देतात आणि प्रवेशयोग्यता



सुधारण्यास मोठ्या प्रमाणात मदत करतात.

- **शिकण्याची क्षमता** : प्रगत नैसर्गिक भाषेचे संस्करण कालांतराने त्यांचे कार्यप्रदर्शन सतत सुधारण्यासाठी मशीन लर्निंगचा वापर करतात. ते मागील संभाषणांमधून शिकतात, त्यांची समज सुधारतात आणि चांगले प्रतिसाद निर्माण करतात.

### मराठी भाषेत NLP च्या वापरासाठी क्षेत्रे :

आज जगभरात इंग्रजी, फ्रेंच, जपानी, कोरियन, जर्मन इत्यादी प्रमुख भाषांमध्ये NLP व AI वर आधारित मोठ्या प्रमाणावर संशोधन व विकासकार्य झाले आहे. मात्र, मराठीसारख्या प्रादेशिक भाषांमध्ये या विषयावरचे संशोधन तुलनेने मर्यादित आहे. मराठी ही भारतातील एक प्राचीन व समृद्ध भाषा असून तिचा वापर करणाऱ्यांची संख्या कोट्यवधींच्या घरात आहे. अशा परिस्थितीत, मराठी भाषेसाठी NLP आणि AI चे उपयोजन अत्यंत महत्त्वाचे ठरते. मराठी भाषेत NLP च्या वापरासाठी अनेक क्षेत्रे असून त्यातून मोठ्या प्रमाणावर उपयोगिता साधता येऊ शकते. उदाहरणार्थ, Speech-to-Text या प्रणालीद्वारे व्याख्याने, मुलाखती, चर्चासत्रे, वृत्तपत्र किंवा बातम्यांचे ऑडिओ स्वरूप सहजपणे मजकुरात रूपांतरित करता येते. हे शिक्षक, विद्यार्थी, पत्रकार किंवा संशोधकांसाठी अतिशय उपयुक्त ठरते. याउलट, Text-to-Speech प्रणालीमुळे एखाद्या मजकुराचा स्वयंचलितपणे आवाज निर्माण केला जातो. विशेषतः दृष्टीदोष असलेल्या नागरिकांसाठी शैक्षणिक व माहितीपर साहित्य ऐकण्यायोग्य स्वरूपात उपलब्ध करणे शक्य होते. तसेच, मशीन अनुवाद (Machine Translation) हे देखील एक अत्यंत उपयोगी तंत्रज्ञान आहे. यामार्फत इंग्रजी ते मराठी, हिंदी ते मराठी किंवा अन्य भाषांमधून मराठीत सहज भाषांतर करता येते. शिक्षण, प्रशासन, न्यायव्यवस्था, ग्राहकसेवा यांसारख्या विविध क्षेत्रात हे अनुवाद अत्यंत उपयुक्त ठरतात. कृत्रिम बुद्धिमत्तेचा आणखी एक महत्त्वाचा भाग म्हणजे भावना विश्लेषण (Sentiment Analysis). यामार्फत सोशल मिडियावर व्यक्त होणाऱ्या प्रतिक्रिया, भावना, ट्रेंड्स यांचा अभ्यास करून जनमताचा अंदाज घेता येतो. नवीन युगात चॅटबॉट्स (Chatbots) आणि व्हर्चुअल सहाय्यक (Virtual Assistants) यांचे महत्त्वही लक्षणीय वाढले आहे. मराठी भाषेतून सेवा देणारे चॅटबॉट्स तयार झाल्यास ग्राहकसेवा, शासकीय माहिती, शैक्षणिक मार्गदर्शन आदी कामांमध्ये मोठी सोय होऊ शकते. उदाहरणार्थ, शेतकऱ्यांना मराठीतून हवामान, खतांची माहिती, सरकारी योजना समजावणारे व्हर्चुअल सहाय्यक तयार होऊ शकतात. तसेच, Optical Character Recognition (OCR) या तंत्रज्ञानाद्वारे छापील किंवा हस्तलिखित मराठी मजकुराचे डिजिटल रूपांतरण करता येते. जुनी पुस्तके, दस्तऐवज, ऐतिहासिक माहिती, ग्रंथ हे डिजिटल स्वरूपात संग्रहित करणे व सर्चयोग्य बनवणे हे OCR च्या साहाय्याने शक्य आहे. मराठी भाषेसाठी NLP आणि AI या क्षेत्रात काम करताना संवेदनशीलता व सांस्कृतिक सुसंगती यांचा विचार करणेही तितकंच महत्त्वाचं ठरतं. कोणतीही भाषा ही केवळ संवादाचं माध्यम नसून ती एक सांस्कृतिक वाहक देखील असते. त्यामुळे NLP प्रणाली तयार करताना त्या भाषेतील पारंपरिक शब्दप्रयोग, सामाजिक संकेत, बोलींचा संदर्भ, म्हणी, समास, लोकोक्ती, आणि धार्मिक भाषिक सत्त्व यांचा विचार केला गेला पाहिजे. उदा. "जन्मठेप", "दगडाच्या काळजाचा", "देव पाण्यात गेला" अशा वाक्यरचनांचे शब्दशः भाषांतर किंवा विश्लेषण केल्यास चुकीचा अर्थ लागू शकतो. म्हणून सांस्कृतिक संदर्भात्मक NLP (Context-Aware NLP) ही एक नवी उपशाखा म्हणून



मराठीसारख्या भाषांसाठी उदयास येऊ शकते. मराठीतील बालसाहित्य, साहित्यिक ग्रंथ, धार्मिक ग्रंथ, विज्ञान व तांत्रिक लेखनाचे डिजिटायझेशन हेही NLP मध्ये मोठ्या संधीचे क्षेत्र आहे. OCR (Optical Character Recognition) द्वारे जुने ग्रंथ, हस्तलिखितं, वृत्तपत्रं यांचे डिजिटायझेशन केल्यास मराठी भाषेचा ज्ञानकोश समृद्ध होऊ शकतो. तसेच, मराठी भाषेचा विकास केवळ शैक्षणिक स्तरापुरता मर्यादित न ठेवता तो समाजाभिमुख करणे आवश्यक आहे. याशिवाय, शालेय व महाविद्यालयीन अभ्यासक्रमात NLP व AI यांचा समावेश करून विद्यार्थ्यांना भाषावैज्ञानिक तंत्रज्ञानाची ओळख करून देणे आवश्यक ठरते. मराठी माध्यमातील विद्यार्थ्यांना या क्षेत्रात करिअर घडवता यावे यासाठी भाषिक अभियांत्रिकी (Computational Linguistics) सारख्या अभ्यासक्रमांची रचना मराठीतूनही व्हावी लागेल. AI नैतिकता (Ethics of AI) ही देखील एक महत्त्वाची बाब आहे. NLP प्रणाली तयार करताना भाषिक पूर्वग्रह, समाजातील संवेदनशील मुद्दे, जाती, लिंग, वर्ग यासंबंधी योग्य न्याय्यता राखली गेली पाहिजे. उदाहरणार्थ, जर भावना विश्लेषण करताना NLP प्रणालीने विशिष्ट जातीच्या लोकांबद्दल नकारात्मक पूर्वग्रह दर्शवला, तर त्यामुळे तांत्रिक प्रणालीवर अवलंबून असलेल्या लोकांवर अन्याय होऊ शकतो. म्हणूनच मराठीसह सर्व भाषांसाठी NLP प्रणाली नैतिकदृष्ट्या पारदर्शक व सामाजिकदृष्ट्या जबाबदार असाव्यात.

### समारोप :

वरील सर्व शक्यतांबरोबरच मराठी भाषेसाठी NLP आणि AI वापरताना अनेक मर्यादा व आव्हाने देखील समोर येतात. प्रमुख अडचण म्हणजे भाषिक डेटाचा अभाव. इंग्रजीसाठी लाखो कोटी मजकूर, दस्तऐवज, ऑडिओ डेटा सहज उपलब्ध आहे; पण मराठीसाठी असे मोठे डेटासेट कमी प्रमाणात उपलब्ध आहेत. यामुळे मॉडेल्स प्रशिक्षित करण्यास मर्यादा येतात. दुसरे मोठे आव्हान म्हणजे बोलींची विविधता. मराठीत विविध प्रादेशिक बोली (जसे की वऱ्हाडी, कोकणी, माळवणी, पुणेरी, खान्देशी इ.) आढळतात. या बोलीत उच्चार, शब्दप्रयोग, व्याकरण आदींचा भेद असल्यामुळे NLP प्रणालींसाठी ही एक जटिल बाब ठरते. शिवाय, व्याकरणातील गुंतागुंत, अनेकवचन, लिंग, काळ, विभक्ती यांची विविध रूपे, वाक्यरचना यामुळे भाषा प्रक्रिया अधिक कठीण होते. शेवटी, तांत्रिक साधनांची कमतरता ही देखील एक गंभीर बाब आहे. इंग्रजीसाठी GPT, BERT, T5 यांसारखी प्रगत भाषा मॉडेल्स विकसित झालेली असली तरी मराठीसाठी असे मॉडेल्स अजून सुरुवातीच्या टप्प्यात आहेत. तसेच, मराठीसाठी वेगळ्या NLP टूल्स, लाइब्रेरीज, APIs यांची आवश्यकता आहे. अशा पार्श्वभूमीवर, मराठी भाषेसाठी NLP आणि AI च्या क्षेत्रात अधिक संशोधनाची, आर्थिक गुंतवणुकीची व धोरणात्मक प्रोत्साहनाची गरज आहे.

### संदर्भ-सूची :

- जोशी, चंद्रशेखर. (२०१९). नैसर्गिक भाषा प्रक्रिया: संकल्पना व उपयोग. सॉफ्टटेक पब्लिकेशन्स, मुंबई.
- पाटील, सुनील. (२०२१). "मराठी भाषेतील बोलीभाषा आणि NLP संदर्भातील अडचणी." भाषा व संगणक, खंड ५, अंक २,



- कुलकर्णी, अनिरुद्ध व भट्टाचार्य, पुष्पा. (२०२०). "भारतीय भाषांसाठी न्युरल मशीन भाषांतर." जर्नल ऑफ कम्प्युटेशनल लिंग्विस्टिक्स इन इंडियन लँग्वेजेस, खंड ८,
- सावरकर, मृणाल. (२०२२). "मराठी द्वीट्सचे भावनिक विश्लेषण: एक अभ्यास." सावित्रीबाई फुले पुणे विद्यापीठातील वार्षिक संशोधन परिषद अहवाल.
- मिश्रा, प्रांजल व सिंह, रोहित. (२०२०). "कमी स्रोत असलेल्या भाषांसाठी NLP साधने विकसित करताना येणाऱ्या अडचणी: मराठीचे उदाहरण." इंटरनॅशनल जर्नल ऑफ लँग्वेज टेक्नॉलॉजी, खंड ३(४),

