विषय	हिंदी
प्रश्नपत्र सं. एवं शीर्षक	**P5: भाषाविज्ञान**
इकाई सं. एवं शीर्षक	**M23: संगणनात्मक भाषाविज्ञान और भाषा-संसाधन**
इकाई टैग	**HND_P5_M23**

निर्माता समूह	
प्रमुख अन्वेषक	**प्रो. गिरीश्वर मिश्र** कुलपति, महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र) 442001 ईमेल : misragirishwar@gmail.com
प्रश्नपत्र समन्वयक	**डॉ. उमाशंकर उपाध्याय** पूर्व प्रोफेसर, भाषा विद्यापीठ महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र) 442001 ईमेल : usupadhyay@gmail.com
इकाई-लेखक	**डॉ. धनजी प्रसाद** असिस्टेंट प्रोफेसर, भाषाविज्ञान एवं भाषा प्रौद्योगिकी विभाग, भाषा विद्यापीठ महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र) 442001 ईमेल : dhpr.langtech@gmail.com
इकाई समीक्षक	**डॉ. उमाशंकर उपाध्याय** पूर्व प्रोफेसर, भाषा विद्यापीठ महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र) 442001 ईमेल : usupadhyay@gmail.com
भाषा संपादक	**डॉ. उमाशंकर उपाध्याय** पूर्व प्रोफेसर, भाषा विद्यापीठ महात्मा गांधी अंतरराष्ट्रीय हिंदी विश्वविद्यालय, वर्धा (महाराष्ट्र) 442001 ईमेल : usupadhyay@gmail.com

पाठ का प्रारूप

1. पाठ का उद्देश्य
2. प्रस्तावना
3. संगणनात्मक भाषाविज्ञान : अवधारणा
4. संगणनात्मक भाषाविज्ञान : विषयवस्तु
5. संगणनात्मक भाषाविज्ञान के अनुप्रयोग क्षेत्र
6. निष्कर्ष

1. पाठ का उद्देश्य

इस पाठ के अध्ययन के उपरांत आप-

- संगणनात्मक भाषाविज्ञान की अवधारणा से परिचित हो सकेंगे।
- इसके विवेच्य विषयों को समझ सकेंगे।
- प्राकृतिक भाषा संसाधन का परिचय प्राप्त कर सकेंगे।
- इसके अनुप्रयोग क्षेत्र के बारे में जान सकेंगे।

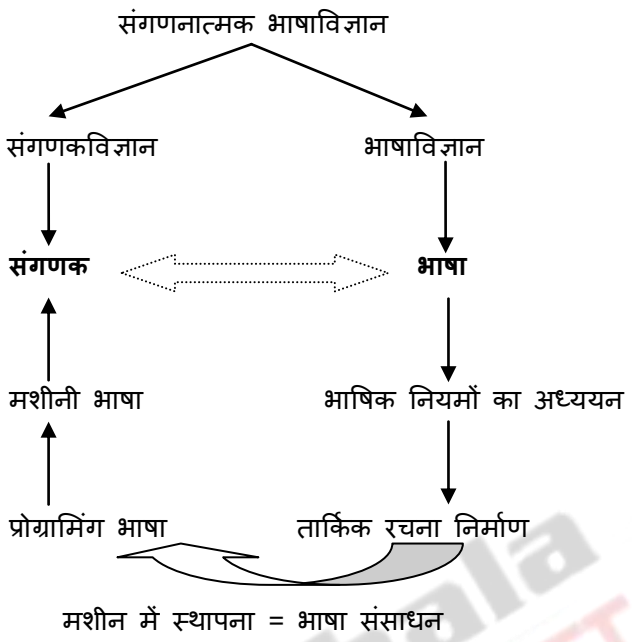
2. प्रस्तावना

संगणक का विकास 20वीं सदी के मध्य में हुआ। इसी समय संगणक में मानव भाषा संबंधी नियमों के अनुप्रयोग और भाषा संबंधी कार्यों को मशीन द्वारा संपन्न कराए जाने के संदर्भ में कुछ पश्चिमी विद्वानों द्वारा काम करना आरंभ किया गया। मूल रूप से यह कार्य मशीनी अनुवाद (Machine translation) से जुड़ा हुआ था। किंतु बाद में अनुप्रयोग क्षेत्रों का विस्तार हुआ, और धीरे-धीरे संगणनात्मक भाषाविज्ञान का उदय हुआ। मानव भाषा को मशीन में स्थापित करने का कार्य भाषा संसाधन है। संगणनात्मक भाषाविज्ञान के अंतर्गत यही कार्य किया जाता है। भाषा संसाधन संबंधी अपने कार्य में अनुप्रयोग क्षेत्रों की बढ़ती माँग के कारण आज संगणनात्मक भाषाविज्ञान असीम संभावनाओं से युक्त ज्ञानक्षेत्र है।

3. संगणनात्मक भाषाविज्ञान की अवधारणा

संगणनात्मक भाषाविज्ञान (Computational linguistics) को संगणकीय भाषाविज्ञान या संगणक भाषाविज्ञान भी कहा गया है। इसमें एक तरफ 'भाषाविज्ञान' (Linguistics) है और दूसरी तरफ 'संगणकविज्ञान' (Computer science) है। भाषाविज्ञान में मानव भाषाओं का अध्ययन-विश्लेषण किया जाता है तथा नियमों की खोज और स्थापना की जाती है। संगणकविज्ञान का केंद्र बिंदु 'संगणक' (Computer) है। संगणनात्मक भाषाविज्ञान में भाषावैज्ञानिक इकाइयों और नियमों को 'संगणक' में स्थापित करने का प्रयास किया जाता है। इस कारण संगणनात्मक भाषाविज्ञान में पूरे संगणकविज्ञान को समझने या जानने की आवश्यकता नहीं होती, बल्कि संगणक की केवल उतनी प्रक्रियाओं को समझने की आवश्यकता होती है, जिनसे भाषा संसाधन संबंधी कार्य किए जा सके। मशीन की अपनी भाषा होती है, जिसे द्विआधारी भाषा (Binary language) कहते हैं। इसे समझना और याद रखना मानव के लिए असंभव जैसा कठिन कार्य है, इसलिए बीच की भाषा 'प्रोग्रामिंग भाषा' का प्रयोग करना पड़ता है। प्रोग्रामिंग भाषा में भाषिक नियमों का प्रयोग करने के लिए उन्हें तार्किक रूप (Logical form) में ढालने की आवश्यकता पड़ती है।

मानव भाषा को मशीन में इस प्रकार से स्थापित करना कि मशीन द्वारा भाषा संबंधी कार्य कराए जा सकें, भाषा संसाधन है। संगणनात्मक भाषाविज्ञान का यही लक्ष्य और कार्य है। इसे निम्नलिखित आरेख द्वारा समझ सकते हैं-



अत: संगणनात्मक भाषाविज्ञान एक ऐसा ज्ञानक्षेत्र है जिसके द्वारा मानव भाषाओं के ज्ञान को मशीन में संसाधित किया जाता है। इस कारण यह 'भाषाविज्ञान' का एक अनुप्रयुक्त क्षेत्र है। इसमें भाषा संबंधी सैद्धांतिक ज्ञान का संगणक में अनुप्रयोग किया जाता है। इसके लिए भाषा के सभी स्तरों (स्वन (ध्वनि), स्वनिम, रूपिम, शब्द, पदबंध, वाक्य और प्रोक्ति) संबंधी सैद्धांतिक ज्ञान की आवश्यकता पड़ती है।

4. संगणनात्मक भाषाविज्ञान : विषयवस्तु

संगणनात्मक भाषाविज्ञान की मूल विषयवस्तु 'प्राकृतिक भाषा संसाधन' (Natural Language Processing : NLP) है। इसे संक्षेप में 'भाषा संसाधन' भी कहा जाता है। इसका कारण यह है कि 'भाषा' शब्द का मूल अर्थ 'मानव भाषा' ही है, और इसे ही संगणकवैज्ञानिकों द्वारा 'प्राकृतिक भाषा' कहा गया है। इसलिए 'प्राकृतिक भाषा संसाधन' और 'भाषा संसाधन' दो चीजें न होकर एक ही हैं। प्राकृतिक भाषा संसाधन मानव भाषाओं के ज्ञान को मशीन में स्थापित करने के लिए किए जाने वाले विभिन्न प्रकार के उपक्रमों का समेकित नाम है। इसके अंतर्गत प्राकृतिक भाषाओं से संबंधित सॉफ्टवेयरों और अनुप्रयोग प्रणालियों के विकास की संपूर्ण प्रक्रिया आ जाती है। इसमें एक तरफ प्राकृतिक (मानव) भाषाओं के अध्ययन का पक्ष है, तो दूसरी तरफ उनके मशीन में अनुप्रयोग (संसाधन) का।

एक क्षेत्र के रूप में प्राकृतिक भाषा संसाधन को परिभाषित करते हुए सामान्य शब्दों में कहा जा सकता है कि 'प्राकृतिक भाषा संसाधन वह क्षेत्र है जिसमें मशीन (कंप्यूटर) और मानव (प्राकृतिक) भाषाओं के बीच अंत:संबंध स्थापित किया जाता है'। अपने व्यापक स्वरूप में 'प्राकृतिक भाषा संसाधन' वह प्रक्रिया है जो किसी मशीन (मुख्यतः कंप्यूटर) को प्राकृतिक भाषा को समझने (to understand), विश्लेषण करने (to analyse), क्रियान्वित करने (to manipulate) और प्रजनन करने (to generate) में सक्षम बनाती है"। यदि मशीन या संगणक में एक बार भाषिक

ज्ञान को स्थापित कर दिया जाए तो प्रयोगकर्ता अपने उद्देश्य के अनुसार विभिन्न प्रकार की अनुप्रयोग प्रणालियों (Application softwares) का विकास या उपयोग कर सकता है।

प्राकृतिक भाषा संसाधन का एक दूसरा पक्ष 'प्राकृतिक भाषा बोधन' (Natural Language Understanding : NLU) भी है। विकिपेडिया (Wikipedia) में प्राकृतिक भाषा बोधन के संबंध में कहा गया है, "Natural Language Understanding is a subtopic of NLP that deals with machine reading comprehension." यह एक प्रकार से प्राकृतिक भाषा संसाधन का एक उपभेद है, जो मशीनी अधिगम से संबंधित है। प्राकृतिक भाषा संसाधन मूलत: मानव भाषाओं को समझने या उनसे संबंधित कार्यों को संपादित करने में मशीन की दक्षता से संबंधित कार्य है। इसमें यह देखा जाता है कि संबंधित भाषा के कार्यों को स्वचलित रूप से संपादित करने में मशीन कितनी सक्षम हो सकी है। अर्थात व्यापक रूप से कहा जा सकता है कि जो कार्य एक संगणनात्मक भाषावैज्ञानिक या प्रोग्रामर की दृष्टि से प्राकृतिक भाषा संसाधन है उसका अगला चरण प्राकृतिक भाषा बोधन है। इसका तात्पर्य यह है कि हम मशीन में जितने प्रकार की संरचनाओं के नियम स्थापित कर देते हैं मशीन उतने प्रकार की भाषिक अभिव्यक्तियों का बोधन और विश्लेषण करने में सक्षम हो जाती है।

मानव व्यवहार में भाषा का प्रयोग दो रूपों में किया जाता है- वाचिक और लिखित। इस दृष्टि से संगणनात्मक भाषाविज्ञान में भाषा संसाधन संबंधी कार्य दो रूपों में किया जाता है- वाक् संसाधन और पाठ संसाधन। इन दोनों को संक्षेप में इस प्रकार समझा जा सकता है-

4.1 वाक् संसाधन (Speech Processing) :

वाक् संसाधन से तात्पर्य भाषा के वाचिक रूप के संसाधन से है। भले ही पहले वाचिक सामग्री को संग्रहीत करना संभव नहीं था, किंतु अब तकनीकी जगत इस स्तर तक उन्नति कर चुका है कि न केवल वाचिक सामग्री को संग्रहीत किया जा सकता है, बल्कि इसे विश्लेषित और संसाधित भी किया जा सकता है। इस दृष्टि से वाक् के क्षेत्र में संगणात्मक भाषाविज्ञान में निम्नलिखित तीन प्रकार के कार्य किए जा रहे हैं-

(क) वाक् अभिज्ञान (Voice Recognition) : इसका संबंध किसी वाचिक सामग्री को एक बार रिकार्ड तथा संग्रहीत करके पुनः उसका प्रयोग उसी प्रकार की ध्वनियों या वाचिक सामग्री से बार-बार मिलान करने से है। उदाहरण के लिए किसी मशीनी प्रणाली (लैपटॉप, मोबाइल आदि) में किसी की आवाज का कूटसंकेत (Password) डाल दिया जाता है। फिर मशीन उसे संग्रहीत कर लेती है, तथा उस मशीन को कभी भी पुनः आरंभ करने के लिए पूर्व में संग्रहीत आवाज से मिलान किया जाता है।

(ख) वाक् विश्लेषण (Speech Analysis) : जब इनपुट के रूप में कोई वाचिक सामग्री मिलती है, तो उसमें वाक्य, शब्द और अक्षर आदि को प्राप्त करना वाक् विश्लेषण है। इसका उपयोग वाक् से पाठ (Speech to Text- STT) में किया जाता है, जहाँ इनपुट वाक् को पाठ में बदलने का कार्य किया जाता है।

(ग) वाक् संश्लेषण (Speech Synthesis) : इसका संबंध छोटी-छोटी वाक् इकाइयों या वाक् खंडों को मिलाकर बड़ी इकाई बनाने से है। इसका उपयोग पाठ से वाक् (Text to Speech- TTS) में किया जाता है।

वाक् संसाधन प्राकृतिक भाषा संसाधन का एक चुनौतीपूर्ण कार्य है, क्योंकि वाचिक भाषा के रूप में ध्वनि तरंगों का संचय होता है। ध्वनि तरंगों में बहुत कम सूचनाएँ प्राप्त करने योग्य होती हैं। इसके अलावा इनके खंडीकरण के

आधार भी बहुत जटिल होते हैं, प्रत्येक व्यक्ति की आवाज अलग होती है। आवाज व्यक्ति की आयु, लिंग, शारीरिक स्थिति आदि के आधार पर भिन्न हो जाती है। इस कारण ध्वनि विश्लेषण अत्यंत कठिन कार्य है।

4.2 पाठ संसाधन (Text Processing)

पाठ संसाधन का संबंध लिखित सामग्री को मशीन के माध्यम से संसाधित करने से है। लिखित भाषा लिपि चिह्नों पर आधारित होती है। इस कारण इस सामग्री का विश्लेषण और संश्लेषण वाचिक रूप की तुलना में अधिक सरल होता है। इसमें लेखन की इकाइयाँ (वर्ण, शब्द, पदबंध वाक्य) आदि स्पष्ट होती हैं। इस कारण इनके नियम अलग प्रकृति के होते हैं। पाठ संसाधन मुख्यतः लिपिचिह्न आधारित संसाधन है। अतः इसके प्रतिमान और कारक भिन्न होते हैं। सैद्धांतिक रूप से भी भाषावैज्ञानिक चिंतन भले ही अपने-आप को वाचिक भाषा पर केंद्रित बताता हो, किंतु यह मुख्यतः लिखित रूप पर ही आधारित है। भाषा की तीन प्रमुख इकाइयों के आधार पर पाठ संसाधन को निम्नलिखित तीन स्तरों पर देखा जा सकता है-

(क) **शब्द संसाधन (Word Processing)** : जब लिखित पाठ में स्वचलित संचालन संबंधी कार्य केवल शब्द स्तर तक किए जाते हैं, तो इस सीमित स्थिति को शब्द संसाधन कहते हैं। इसमें मुख्य रूप से वर्तनी परीक्षण, फॉन्ट डिजाइनिंग, टेक्स्ट फॉर्मेटिंग आदि आते हैं।

(ख) **वाक्य संसाधन (Sentence Processing)** : इसमें पाठ के अंतर्गत वाक्य स्तर तक मशीन द्वारा अभिज्ञान स्वचलित संसाधन या संचालन की क्षमता होती है। टैगिंग, पदबंध चिह्नन और व्याकरणिक परीक्षण आदि संबंधी कार्य इसमें आते हैं।

(ग) **पाठ संसाधन (Text Processing)** : इसमें मशीन के अंतर्गत पूरे पाठ को समझते हुए उसे संसाधित करने की क्षमता होती है। यह पाठ (या प्रोक्ति) और अर्थ आदि सभी को अपने अंदर समाहित करता है। मशीनी अनुवाद और कृत्रिम बुद्धि के क्षेत्र में किए जा रहे कार्य इसी स्तर के होते हैं।

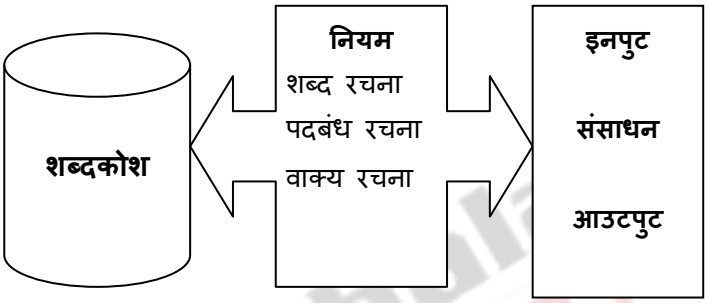
पाठ संसाधन के उपर्युक्त स्तर भिन्न-भिन्न न होकर समावेशी हैं। वाक्य स्तर में शब्द स्तर का और पाठ स्तर में शब्द और वाक्य दोनों स्तरों का समावेश होता है। चूँकि व्याकरणिक दृष्टि से वाक्य सबसे बड़ी इकाई है, इस कारण सामान्यतः पारंपरिक रूप से वाक्य संसाधन को भी 'पाठ संसाधन' के अंतर्गत देखा जाता रहा है। किंतु वर्तमान में पाठ के स्तर पर कई बड़े कार्य हुए हैं, जिनमें अन्वादेश (Anaphora), पूर्वादेश (Cataphora) तथा संदर्भ अभिव्यक्तियाँ (Reference expressions) जैसी रचनाओं के विश्लेषण से एकाधिक वाक्यों में संबंध और संदर्भ देखने की आवश्यकता महसूस हुई है। इस कारण अब पाठ स्तर का संसाधन और वाक्य स्तर के संसाधन से अधिक व्यापक बनता जा रहा है।

वाक् संसाधन और पाठ संसाधन के संदर्भ में एक महत्वपूर्ण बात यह है कि वाक् संसाधन की केंद्रीय इकाई 'ध्वनि तरंग' (Sound wave) है। इसलिए ध्वनि तरंगों के विश्लेषण और विवेचन से जुड़ी इकाइयाँ और प्रविधियाँ वाक् संसाधन में आती हैं। पाठ लिखित सामग्री से संबद्ध है, किंतु इसमें केवल लिपि चिह्नों का विश्लेषण ही नहीं आता, बल्कि शाब्दिक और वाक्यात्मक गठन संबंधी विश्लेषण भी इसी में आता है। ध्वनि तरंगों का विश्लेषण करके पदबंध और रूपिम जैसी इकाइयाँ नहीं प्राप्त की जा सकतीं। इसलिए इस प्रकार का विवेचन और विश्लेषण पाठ संसाधन का अंग होता है।

प्राकृतिक भाषा संसाधन के उपागम (Approaches to NLP)

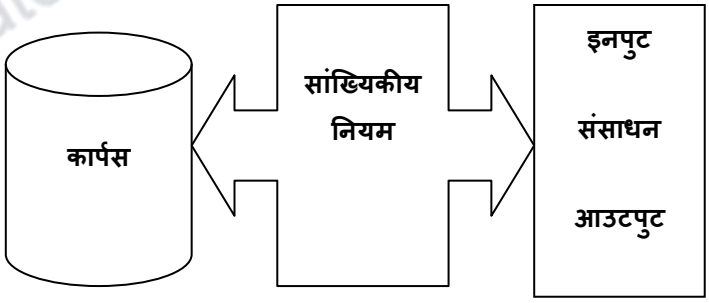
संगणनात्मक भाषाविज्ञान में भाषा संसाधन संबंधी कार्य दो विधियों से किया जाता है- नियम आधारित और कार्पस आधारित। इन्हें प्राकृतिक भाषा संसाधन के दो उपागम कहा जाता है। इन्हें संक्षेप में इस प्रकार देखा जा सकता है-

1. **नियम आधारित उपागम (Rule-based approach)** : यह अध्ययन भाषावैज्ञानिक पृष्ठभूमि पर केंद्रित है। इसमें 'नियम' से तात्पर्य भाषावैज्ञानिक नियमों से है। इसमें भाषिक इकाइयों के संग्रह के रूप में केंद्रीय घटक 'शब्दकोश' (Lexicon) होता है। इसके उपरांत रूपवैज्ञानिक (Morphological), पदबंधीय (Phrasal) तथा वाक्यात्मक (Syntactic) नियम तार्किक रूप से दिए जाते हैं। उनके आधार पर ही इनपुट सामग्री पर कार्य किया जाता है। इसे चित्र के रूप में इस प्रकार देखा जा सकता है-



इसमें देखा जा सकता है कि संसाधन के लिए 'शब्दकोश + नियम' विधि का इस उपागम हेतु प्रयोग किया जाता है।

2. **कार्पस आधारित उपागम (Corpus-based approach)** : कार्पस आधारित उपागम में कार्पस मूल भाषावैज्ञानिक स्रोत होता है। कार्पस किसी भाषा की वाचिक या लिखित (या दोनों) प्रकार की सामग्री ऐसा विशाल संग्रह है जो प्रातिनिधिक (Representative) और मशीन पठनीय (Machine readable) होता है| इसमें भाषा व्यवहार के सभी क्षेत्रों से पाठों का संग्रह किया जाता है। यह सामग्री इतनी विशाल और वैविध्यपूर्ण होती है कि इसमें भाषा की लगभग सभी प्रकार की रचनाएँ आ जाती हैं। अब इस सामग्री पर सांख्यिकीय नियमों (Statistical rules) का प्रयोग किया जाता है, जिससे मशीन इनपुट को संसाधित करके आउटपुट प्रदान करती है। इसी कारण भाषा संसाधन के इस उपागम को सांख्यिकीय उपागम (Statistical approach) भी कहा गया है। इसे निम्नलिखित प्रकार से देख सकते हैं-



कार्पस आधारित उपागम का प्रयोग करके किए जाने वाले भाषा संसाधन को कार्पस आधारित भाषा संसाधन कहते हैं। मशीनी संसाधन के लिए केवल विशाल डाटा संग्रह के रूप में कार्पस बहुत उपयोगी नहीं होते, बल्कि उनमें शब्दों के साथ विभिन्न प्रकार की भाषिक (स्वनिक, व्याकरणिक और आर्थी) सूचनाएँ भी देनी पड़ती हैं। उनके शब्दवर्ग टैग की सूचना प्रदान करना इनमें से पहला आधारभूत कार्य है जिसे कार्पस की शब्दभेद टैगिंग (या संक्षेप में कार्पस टैगिंग) कहते हैं। टैग संबंधी सूचना या किसी भी प्रकार की सूचनाओं को शब्द, वाक्य या पाठ के साथ जोड़ना 'एनोटेशन' कहलाता है। जोड़ी जाने वाली सूचना की प्रकृति और उसके भाषिक स्वरूप के आधार पर एनोटेशन के विविध प्रकार भी होते हैं, जैसे- रूपवैज्ञानिक एनोटेशन, वाक्यात्मक एनोटेशन, आर्थी एनोटेशन और शैली संबंधी एनोटेशन आदि।

सामान्य कार्पस की जगह एनोटेटेड कार्पस ही भाषा संसाधन की दृष्टि से अधिक उपयोगी होता है।

5. संगणनात्मक भाषाविज्ञान के अनुप्रयोग क्षेत्र

5.1 मशीनी अनुवाद (Machine Translation)

एक भाषा (भा$_1$) के पाठ का दूसरी भाषा (भा$_2$) के पाठ में मशीन के माध्यम से किया जाने वाला रूपांतरण मशीनी अनुवाद है। इसमें मशीन का तात्पर्य मुख्यत: संगणक से है। मशीनी अनुवाद के लिए स्रोत भाषा और लक्ष्य भाषा दोनों के ही वाक्यात्मक नियमों और शब्दकोश की आवश्यकता होती है। वाक्यात्मक नियमों को एल्गोरिद्मों एवं तार्किक अभिव्यक्तियों तथा शब्दकोश को डाटाबेस के रूप में मशीन में संग्रहीत करते हुए मशीनी अनुवाद प्रणालियों (MT Systems) के विकास का कार्य किया जाता है। इस प्रकार से विकसित प्रणालियों को नियम आधारित मशीनी अनुवाद प्रणाली कहते हैं। इन प्रणालियों द्वारा पूर्णत: स्वचलित या आंशिक स्वचलित रूप से अनुवाद का कार्य किया जाता है। मशीनी अनुवाद प्रणालियों के विकास की दूसरी विधि सांख्यिकीय (Statistical) या कार्पस आधारित भी है।

5.2 प्राकाशिक अक्षर संज्ञान (Optical Character Recognition)

प्रकाशिक अक्षर संज्ञान किसी मुद्रित (Printed), टंकित (Typed), या हस्तलिखित (Handwritten) पाठ के स्कैन किए हुए चित्र (Image) को मशीन पठनीय पाठ (Machine readable text) में बदलने की प्रक्रिया है। इसमें सर्वप्रथम किसी पाठ; जैसे : पुस्तक, पत्रिका, समाचार-पत्र आदि के पृष्ठ को स्कैनर की सहायता से स्कैन किया जाता है। इसके पश्चात् स्कैन किए हुए पाठ चित्र (Text image) के प्रत्येक वर्ण को मशीनी पाठ के वर्ण की तरह चिह्नित किया जाता है जिससे आवश्यकता पड़ने पर उसे परिवर्तित और परिवर्धित किया जा सके।

5.3 पाठ से वाक् और वाक् से पाठ (Text to Speech and Speech to Text)

पाठ से वाक् (TTS) प्रणालियों द्वारा 'पाठ' के रूप में दिए गए इनपुट को उच्चरित या वाचिक रूप में बदल दिया जाता है। इसके अंतर्गत मुख्यत: तीन प्रक्रियाएँ होती हैं- पाठ विश्लेषण, लेखिम (Grapheme) से स्वनिम (Phoneme) रूपांतरण और वाक् संश्लेषण। किंतु इन प्रक्रियाओं को संपन्न करने में पाठ से वाक् प्रणाली कई चरणों से होकर गुजरती है। पाठ विश्लेषण और वाक् संश्लेषण की दृष्टि से मुख्यत: चार स्तरों पर इन प्रणालियों को विकसित किया जाता है : अक्षर, शब्द, पदबंध एवं वाक्य स्तर। चारों ही स्तरों पर विकसित की गई प्रणालियों की अपनी उपयोगिता एवं सीमा होती है। किसी भी पाठ से वाक् प्रणाली की गुणवत्ता और उपयोगिता के मापन के तीन आधार बताए गए हैं : सृजनात्मकता (Generativity), शुद्धता (Accuracy) एवं सहजता या स्वाभाविकता (Naturalness)।

वाक् से पाठ (STT) में बोलकर (वाचिक भाषा में) इनपुट दिया जाता है जिसका प्रणाली द्वारा संसाधन किया जाता है और उसे लिखित पाठ के रूप में परिवर्तित कर दिया जाता है। इसमें मुख्यत: चार प्रक्रियाएँ होती हैं- वाक् संज्ञान, वाक् विश्लेषण, वाक् से पाठ रूपांतरण और पाठ संश्लेषण। भाषा के वाचिक रूप का किसी भी स्वचलित प्रणाली द्वारा ठीक-ठीक संज्ञान और विश्लेषण करना बहुत ही जटिल कार्य होता है, क्योंकि वाचिक भाषा का इनपुट ध्वनि तरंगों (Sound waves) के रूप में प्राप्त होता है जिनके आयाम (Amplitude) और आवृत्ति (Frequency) आदि विशेषताओं का विश्लेषण करके स्वनिमों की पहचान करनी होती है। यह बहुत जटिल कार्य है। इसके अतिरिक्त शब्द और वाक्य के चिह्न के लिए आवश्यक वास्तविक या संभावित विराम (Actual or potential pauses) प्राय: प्राप्त नहीं होते। इसी कारण वाक् से पाठ प्रणालियों के विकास कार्य को अत्यंत चुनौतीपूर्ण माना गया है।

5.4 सूचना प्रत्यानयन/पुन:प्राप्ति (Information Retrival- IR) : किसी डाटा संग्रह में संक्षिप्त सूचना के आधार पर केवल संबंधित सूचनाओं या कथनों को प्राप्त करना सूचना प्रत्यानयन है। इंटरनेट पर सूचनाओं का बहुत विशाल भंडार है। साथ ही इसमें रखी हुई सूचनाएँ असंरचित (Unstructured) या बिखरी हुई हैं जो विश्व के किसी भी कोने में और किसी भी रूप में हो सकती हैं। सूचनाओं की इतनी विशाल मात्रा और असंरचित स्वरूप में से केवल उपयोगी सूचनाओं को खोजकर लाना बहुत ही कठिन कार्य है। यह कार्य 'खोज इंजनों' (Search engines) द्वारा किया जाता है जिनमें विविध प्रकार की सूचना पुन:प्राप्ति प्रणालियाँ लगी रहती हैं।

5.5 स्वचलित पाठ सारांशीकरण (Automatic Text Summarization)

किसी बड़े पाठ के मूल अर्थ को बनाए रखते हुए सीमित शब्दों में पुन: प्रस्तुतीकरण 'पाठ सारांशीकरण' है। यह कार्य जब किसी संगणक प्रणाली द्वारा स्वचालित रूप से किया जाता है तो इसे 'स्वचालित पाठ सारांशीकरण' (Automatic text summarization) कहते हैं। पाठ सांराश के निर्माण में इस बात का ध्यान रखा जाता है कि सारांश के रूप में निर्मित पाठ में मूल पाठ की सभी महत्वपूर्ण बातें आ जाएँ। अत: पाठ सारांशक में महत्वपूर्ण बिंदुओं को पहचानने की क्षमता का विकास करना आवश्यक होता है।

5.6 संगणकीय कोश (Computational Lexicon)

मुद्रित कोशों के निर्माण, अध्ययन एवं विश्लेषण की अपनी सीमाएँ रही हैं। संगणकीय कोशों (शब्दवृत्तों) का निर्माण अंकीय रूप (Digital form) में किया जाता है। एक बार अंकीय रूप में कोश का निर्माण हो जाने के पश्चात् उसका प्रयोग करना या उसमें सुधार करना अत्यंत ही सरल कार्य होता है, क्योंकि संगणकीय कोशों में खोज करने (Searching), क्रमानुसार करने (Sorting), मिटाने (Deleting), और सुधारने (Modifying) से संबंधित कार्यों के लिए अलग से सुविधा दी गई रहती है।

5.7 संगणक साधित भाषा अधिगम (Computer Assisted Language Learning- CALL)

इसका संबंध भाषा अधिगम (और भाषा शिक्षण) के क्षेत्र में संगणक के अनुप्रयोग से है। संगणक साधित भाषा अधिगम में शिक्षार्थियों के लिए शिक्षार्थी केंद्रित पाठ (Student centered lessons) निर्मित किए जाते हैं जो दो प्रकार के होते हैं- एकदिशात्मक या व्यक्तिगत अधिगम (Unidirectional or Individualized learning) तथा द्विदिशात्मक या अंतरक्रियात्मक अधिगम (Bidirectional or Interactive learning)। एकदिशात्मक अधिगम में शिक्षार्थी संगणक से केवल सामग्री और निर्देश प्राप्त कर सकता है जबकि द्विदिशात्मक अधिगम में फीडबैक एवं अंतरक्रिया की भी सुविधा रहती है।

5.8 कृत्रिम बुद्धि (Artificial Intelligence - AI)

कृत्रिम बुद्धि एक बहुत बड़ा अनुप्रयुक्त क्षेत्र है जिसका एक भाग 'प्राकृतिक भाषा संसाधन' है। यह एक बृहद ज्ञानक्षेत्र है जिसमें मशीनों में तर्क करने, सोचने या विचार करने, कार्य करने तथा निर्णय लेने की क्षमताओं का विकास किया जा रहा है। इस प्रकार की मशीनों को 'बुद्धिमान एजेंट' (Intelligent agent) भी कहा गया है। बुद्धिमान एजेंट का विकास कृत्रिम बुद्धि का मुख्य उद्देश्य है। इसके लिए मनुष्य को प्रतिरूप (Model) माना गया है। इसमें चित्र संसाधन (Image processing), मोटर संचालन, हार्डवेयर डिजाइन और भाषा संसाधन आदि अनेक क्षेत्र एक साथ मिलकर कार्य करते हैं।

6. निष्कर्ष

संक्षेप में संगणनात्मक भाषाविज्ञान एक नवीन किंतु असीम संभावनाओं से भरा क्षेत्र है। इसका मूल कार्य भाषा संसाधन है, जिसके अंतर्गत मानव भाषाओं से संबंधित ज्ञान को मशीन में इस प्रकार से स्थापित करने का प्रयास किया जाता है कि उनसे भाषा संबंधी कार्य कराए जा सकें। इसमें भाषा संसाधन भाषा के दोनों रूपों- वाचिक और लिखित के आधार पर किया जाता है। इससे दो प्रकार के संसाधन क्षेत्र निर्मित होते हैं, जिन्हें वाक् संसाधन और पाठ संसाधन कहते हैं। भाषा संसाधन संबंधी कार्य 'नियम आधारित' तथा 'कार्पस आधारित' दोनों ही उपागमों से किया जाता है। इसके अनेक महत्वपूर्ण अनुप्रयोग क्षेत्र हैं, जिनमें से मशीनी अनुवाद से लेकर कृत्रिम बुद्धि तक आठ के बारे में ऊपर संक्षिप्त परिचय दिया गया है।

e-Pathshala पाठशाला A Gateway to All Post Graduate Courses