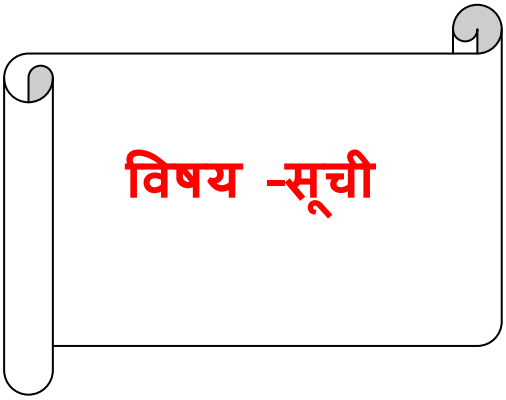
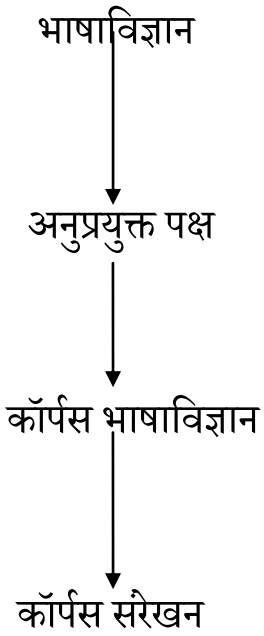


विषय -सूची

भूमिका

कॉर्पस भाषाविज्ञान ,भाषाविज्ञान एवं तकनीकी के बीच से निकली हुई एक अत्यंत तीव्र गति से उतरती एवं विकसित हुई विधा है । कॉर्पस किसी प्राकृतिक भाषा के पाठ (लिखित या मौखिक) का मशीनी संग्रह होता है। इसका अध्ययन कॉर्पस भाषाविज्ञान के अंतर्गत किया जाता है। इसका अध्ययन कॉर्पस भाषाविज्ञान के अंतगर्त किया जाता है । "कॉर्पस भाषाविज्ञान", भाषाविज्ञान का अनुप्रयुक्त पक्ष है।

निम्न आरेख द्वारा हम इसे स्पष्ट रूप से देख सकते हैं:-



संगणकीय भाषाविज्ञान में कॉर्पस एक प्राथमिक स्तोत है। बिना कॉर्पस के भाषा प्रौद्धोगिकी के क्षेत्र में भाषा विषयक कोई भी अनुसंधान संपादित नहीं किया जा सकता है।

संचित ज्ञान(Repository Knowledge) के अंतगर्त भाषाविरण को नियमबद्ध करके कंप्यूटर में फीड किया जाता है , जिसका उपयोग प्राकृतिक भाषा संसाधान के दौरान होता है। किसी भी भाषा के लिखित या वाचिक पाठ रूप को तकनीक के जारिए से जोड़ता है ।

इस शोध में मेरा विषय है **"मणिपुरी से हिंदी के संदर्भ में : कॉर्पस संरेखन"**

इसमें कॉर्पस alignment के लिए data collection एक मणिपुरी उपन्यास किताब है हिंदीतर भाषी हिंदी लेखक पुरस्कार से पुरस्कृत "शीतलजीत की उपन्यास-कृति **माँ** (मणिपुरी उपन्यास **'इमा'**)" और इस का **अनुवादक** है सिजगुरुमयूम ब्रजेश्वर शर्मा । **प्रथम संस्करण** :२०००,**द्वितीय** संस्कारण :२०१२ **प्रकाशन:**अशांबा कम्यूनिकेशन ,इम्फाल द्वारा ,आश्रम पब्लिकेशन्स,इम्फाल के लिए **पता:** ब्रह्मपुर नहाबम पुरितमयूम लैरक , इम्फाल ईस्ट-७९५००१,मणिपुर , भारत ।

A **parallel corpus** is a corpus that contains a collection of original texts in language L_1 and their translations into a set of languages L_2 ... L_n. In most cases, parallel corpora contain data from only two languages.

कंप्यूटर प्रौद्योगिकी के आविष्कार और प्रगति में भाषाविज्ञान में एक नया आयाम जोड़ा और Computational Linguistics नामक एक नए अनुशासन ने जन्म लिया जो Artificial Intelligence का एक महत्वपूर्ण अंग है।

Corpus Linguistics, Computational Linguistics की महत्वपूर्ण शाखा है जो Statistics की प्रविधि का अनुसरण करते हुए अत्यंत व्यवस्थित रूप में भाषा प्रयोग का बड़ी मात्र में प्रत्येक्षित प्रमाण प्रस्तुत करके बहुत महत्वपूर्ण भूमिका निभाती है। साथ ही यह कॉर्पस के विश्लेषण एवं सूचना दोहन (Information Extraction) की सूक्षम युक्तियों में समुच्चय प्रदान होता है , जो मानव भाषा को समझने और मानवविज्ञान के विभिन्न क्षेत्रों में प्रयोग के लिए Computational Linguistics या Artificial Intelligence में आवश्यक माना जाता है।

अंग्रेजी शब्द कॉर्पस Latin शब्द Corpus से व्युतपन्न हुआ है जिसका अर्थ है शरीर i.e. body in English.

कॉर्पस linguistics के संदर्भ में corpus भाषा प्रयोग के किसी विशिष्ट विभेद को प्रस्तुत करने हेतु वैज्ञानिक विधि से संग्रहित मशीन पाठ्य रूप(Machine Readable Form) में उपलब्ध लिखित एवं बोले गए पाठों के नमुने का विशाल संकलन है।

कॉर्पस प्रबंधन के लिए कई उपकरण होते हैं-

१. आवृत्ति गणक (Frequency Counter)
२. KWIC (Key word in context)
३. KWOC (Key word out of context)

४. संरेखन (Aligner/ Alignment)

५. कॉर्पस टैगर (Corpus Tagger)

६. सुसंगतता (Concordance)

साहित्य का पुनरावलोकन :

इस शोध-कार्य को संपन्न करने के लिए विभिन्न सहायक- सामाग्री की सहायता ली गई है।

जिनमें अंग्रेजी , मणिपुरी , हिंदी की पुस्तकें , शोध आलेखों का पुनरावलोकन किया गया है।

जिसका विस्तृत विवरण प्रस्तुत है।

१. En.wikipedia.org/wiki/ParallelCorpus

२. WWW.glottopedia.org/index.php/Parallel Corpus

३. मल्होत्रा , विजय कुमार (2000) कंप्यूटर के भाषिक अनुप्रयोग , वाणी प्रकाशन , नई दिल्ली

५.हिंदी और मणिपुरी दोनों भाषाओं के वाक्य रचना को अच्छी तरह समझना चाहिए क्योंकि दोनों भाषाओं की वाक्य संरचना और वाक्य गठन में कई समानता और कई विषमता मिलती है।

जैसे:-

मणिपुरी:- নুমিৎনা ময়ুম চংখরবা কাল, য়ুম খুদিংগী মৈ-রা হৌগৎনরকপা মতম ।

हिंदी:- सूर्यास्त की वेला थी, घरों में दिया-बाती की वेला।

मणिपुरी:- মরমওকপী লাইনিং-চৎন-খঙলবী য়ুমগোনবীশিংনা সুমাং-থেলোং-ঙাল্লিবী,

বৃন্দাদেবী মথোংদা ধুপ-থাওমৈ থান্দুনা তাইবংশেম্বা মবুঙো শ্রীকৃষ্ণবু নিংশিংন- রকপা

हिंदी :- पूजापाठी गृहलक्ष्मियों आँगन में ,प्रतिष्ठित बृंदादेवी को धूप-दीप अर्पित करते हुए सृष्टिकर्ता भगवान श्रीकृष्ण के ध्यान में लीन थीं।

६.Chafe, W., J. DuBois , and S. Thompson. 1991. Towards a New Corpus of Spoken American English. In K. Aijmer and B. Altenberg (eds) , English Corpus Linguistics. London: Longman.

७. मेरे संजाल में कॉर्पस संरेखन Parallel Corpus Aligner CIIL, Mysore , IIIT Hyderabad, IIIT Guwahati में कार्य जारी है और सब अभी process में हैं।

अध्यायीकरण:-

प्रथम –अध्याय

1.1 कॉर्पस संरेखन: स्वरूप एवं वर्तमान स्थिति

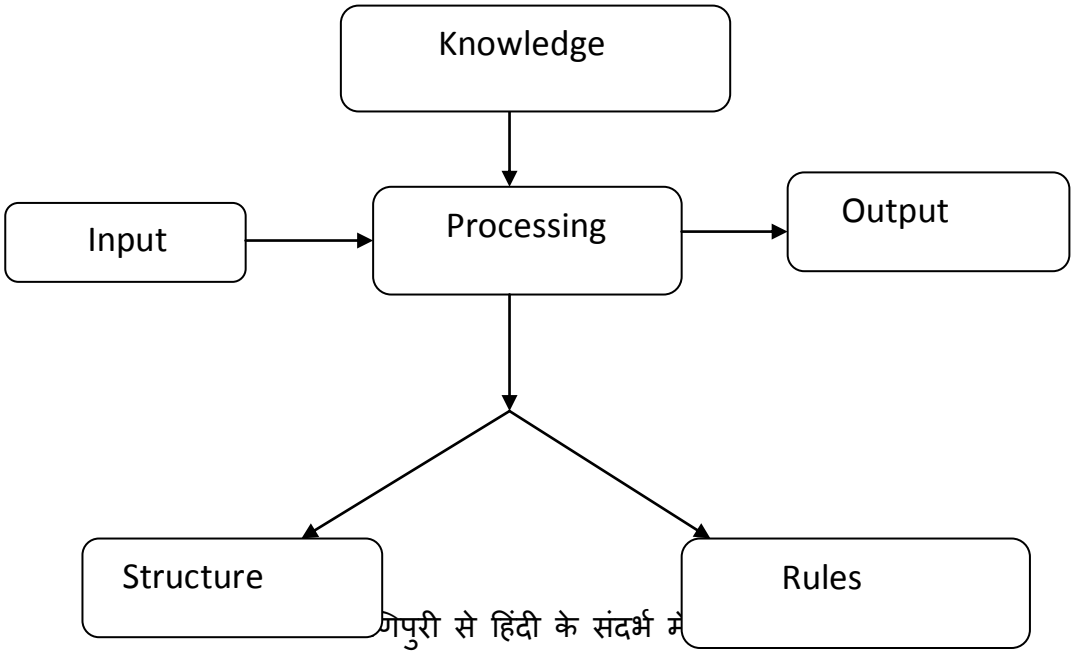
1.2 परिभाषा

1.3 प्रकार

1.4 निर्माण प्रक्रिया

द्वितीय –अध्याय

प्राकृतिक भाषा संसाधन (NLP): प्राकृतिक भाषा मानव-जीवन का भिन्न अंग हैं । मानवों के बीच परस्पर संवाद स्थापित करने और सूचनावों को लिपिबद्ध करके अभ लिखित करने का यह एक प्रमुख साधन है। इसके माध्यम से मानव सूक्षम ,जटिल,गहन और व्यापक विचारों को अभिव्यक्त करने में सफल हो पाता है। वर्तमान युग में कंप्यूटर के अधिकाधिक प्रयोग द्वारा मानव भाषा के ही जैसे मशीनी भाषा के निर्माण के लिए निरंतर प्रयास किए जा रहें है। ताकि मशीन के साथ संवाद संभव हो सके। प्राकृतिक भाषा संसाधन अभिकलनात्मक भाषाविज्ञान का वह अंग है जिसका उद्देश्य प्रत्येक भाषिक इकाई को संसाधित करके ऐसे मॉडल और डिज़ाइन तैयार करना है जिसकी सहायता से मानव-मशीन के बीच संवाद हो सके। मानव मस्तिष्क में भाषा का ज्ञान पर वह व्यवहार करता है। उसी प्रकार कंप्यूटर के अंदर processing के द्वारा memory और भाषा ज्ञान को संपादित किया जाना प्राकृतिक भाषा संसाधन का प्रमुख कार्य क्षेत्र है। इसे आगे के पृष्ट पर आरेख द्वारा स्पष्ट किया गया हैं,



प्राकृतिक भाषा संसाधन में मानव द्वारा लिखित तथा उच्चरित भाषा को कंप्यूटर के माध्यम से संसाधित कराया जाता है। जिसमें ज्ञान ,संरचना तथा नियम आवश्यक घटक के रूप में कार्य करते हैं , लेकिन भाषा में व्याप्त अर्थ की समस्या के कारण प्राकृतिक भाषा संसाधन की प्रक्रिया जटिल होती है।

Natural language processing (**NLP**) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction.

"प्राकृतिक भाषा संसाधन वह क्षेत्र है जिसमें मशीन (कंप्यूटर) और मानव (प्राकृतिक) भाषाओं के बीच अंतरसंबंध स्थापित किया जाता है "।

"प्राकृतिक भाषा संसाधन वह प्रक्रिया है जो किसी मशीन (मुख्यतः कंप्यूटर) को प्राकृतिक भाषा को समझने (understanding) के साथ-साथ विश्लेषण (analysis),क्रियान्वयन(manipulation) और प्रजनन (generation) में सक्षम बनाती है" ।

प्राकृतिक भाषा संसाधन में भाषा के उच्चरित (spoken) और लिखित(written) दोनों ही रूपों पर कार्य किया जाता है। इस प्रक्रिया से विकसित की जाने वाली प्रणालियाँ मानव और मशीन के बीच भाषिक बाधा को तोड़ने वाले अंतरापृष्ट (interface) का कार्य करती हैं। भाषा के उच्चरित और लिखित स्वरूप के संसाधन के दो प्रकार हैं:-

1. वाक संसाधन (Speech Processing):- वाक संसाधन के अंतर्गत भाषा के वाचिक रूप का संसाधन किया जाता है। अर्थात इसमें भाषिक प्रतीक ध्वनि तरंगों (wave sound) के रूप में संचित होते हैं और इनके आधार पर ही विभिन्न अनुप्रयोग प्रणालियों का विकास किया जाता है।

2.पाठ संसाधन (Text processing):-पाठ संसाधन के अंतर्गत भाषा के लिखित रूप का संसाधन किया जाता है। अतः इसके अंतर्गत किए जाने वाले संसाधन कार्य में लेखिम (Grapheme) या लिपि - चिन्हों आधार का कार्य करते हैं। यदि इसे वाक संसाधन के साथ तुलनात्मक रूप से देखा जाए तो ध्वनि तरंगों के रूप में वाक प्रतीकों के विश्लेषण से लिपि-चिन्हों के रूप में भाषिक प्रतीकों के आधार पर पाठपरक भाषिक सामाग्री का विश्लेषण अपेक्षाकृत सरल होता है। इस कारण पाठ संसाधन से जुड़ी प्रणालियों का विकास अधिक किया जाता है।

वास्तव में, प्राकृतिक भाषा संसाधन के अंतर्गत चाहे वाक संसाधन हो या पाठ संसाधन, रुपिमिक और व्याकरणिक नियम एक ही होते हैं केवल स्वनिमिक सामग्री का अंतर होता है।

प्राकृतिक भाषा संसाधन एक अत्यंत व्यापक प्रक्रिया है। किसी भी भाषा के लिए एकल संसाधन इकाई (single processing unit) का विकास अत्यंत जटिल और कठिन कार्य है जो अभी तक सफल नहीं हो सका है। समान्यतः प्राकृतिक भाषा संसाधन प्रणालियों के विकास में दो प्रकार के उपागमों का प्रयोग किया जाता है :

1. **नियम आधारित उपागम (Rule-based Approach):-**

   इस उपागम में व्याकरणिक नियमों को आधार बनाया जाता है। इसमें संबंधित भाषा के व्याकरणिक नियमों को तार्किक रूप (logical form) में संग्रहीत किया जाता है और उसके आधार पर अनुप्रयोग प्रणालियों का विकास किया जाता है। समान्यतः इसमें नियमों के साथ-साथ शब्दकोश की भी आवश्यकता होती है।

   2. संखियकीय उपागम (Statistical Approach):-

   इस उपागम में संबंधित भाषा में प्राप्त प्रामाणिक और वैविध्य से युक्त विशाल डाटा-संग्रह को आधार बनाया जाता है। समान्यतः यह संग्रह कॉर्पस के रूप में किया जाता है। कॉर्पस में संग्रहीत डाटा के आधार पर ही मैपिंग नियमों आदि के माध्यम से इस उपागम के अंतर्गत विभिन्न अनुप्रयोग प्रणालियों का विकास किया जाता है।

प्राकृतिक भाषा संसाधन के कार्य की क्षेत्र(Major tasks in NLP)

१. मशीनी अनुवाद(Machine Translation)

२. वाक अभिज्ञानक(Speech Recognition)

३. वाक संलेशन (Speech Synthesis)

४. सूचना संचयन(Information Extraction)

५. सूचना प्रत्यानयन (Information Retrieval)

६. प्रकाशित संप्रतिक अभिज्ञान(Optical Character Recognition)

७. वर्तनी संशोधक(Spell Checker)

८. एकीकृत शब्द संसाधक(Integrated word Processor)

९. नाम पहचानक(Name Entity Recognition)

१०. संदिग्धार्थी विश्लेषक (Ambiguity Analyzer)

११. कोश निर्माण (Lexicography)

उपायुक्त सभी क्षेत्रों में संगणक की अभिन्न भूमिका है। आज कंप्यूटर की प्रकार्यात्मक क्षमता एवं हमारी अवधारण में निरंतर विस्तार होता जा रहा है,जिसके फलस्वरूप अनेक भाषाई उपकरणों का निर्माण किया जा रहा है।

भाषा प्रौद्योगिकी

भाषा के वैज्ञानिक अध्ययन अनुसंधान की दिशा में भाषाविज्ञान को एक ऐसे ज्ञानानुसर के रूप में ग्रहण किया जा सकता है जो प्रकृति और संस्कृति , जैविकी और सामाजिकी , कारण और कल्पना ,प्रौद्योगिकी और सैद्धांतिकी तथा प्राकृतिक विज्ञान और मानविकी आदि आसमान संरचना वाले विषयानुशासनों के बीच से विकसित हुआ। भाषाविज्ञान के अध्ययन क्षेत्र में तकनीक जुड़ जाने से एक नये विषय भाषा प्रौद्योगिकी का प्रादुर्भाव हुआ। अर्थात भाषाविज्ञान के अध्ययन क्षेत्र में यह एक नया आयाम जुड़ गया है। यह एक अतुल्य-संकल्पना है। भाषा प्रौद्योगिकी एवं भाषाविज्ञान के अंतगर्त अनेवाली विधाओं को

तकनीकी से जोड़कर सूचना प्रौद्योगिकी के क्षेत्र में हो रहे नित-नए विकास को क्रियान्वित करने का प्रयास किया जा रहा है।

भाषा प्रौद्योगिकी की परिभाषा (Definition of Language Technology)

"भाषा द्वारा प्रयास और मानव शरीर की परिधि से बाहर इसके प्रकार्यों के स्वचालन की दृष्टि से तकनीकी का उपयोग तथा तकनीकी द्वारा ज्ञान आधारित स्वचालन के लिए भाषा का उपयोग स्वचालन के लिए भाषा का उपयोग 'भाषा तकनीकी' है।

भाषा प्रौद्योगिकी की परिभाषा इस प्रकार दी जा सकती है-

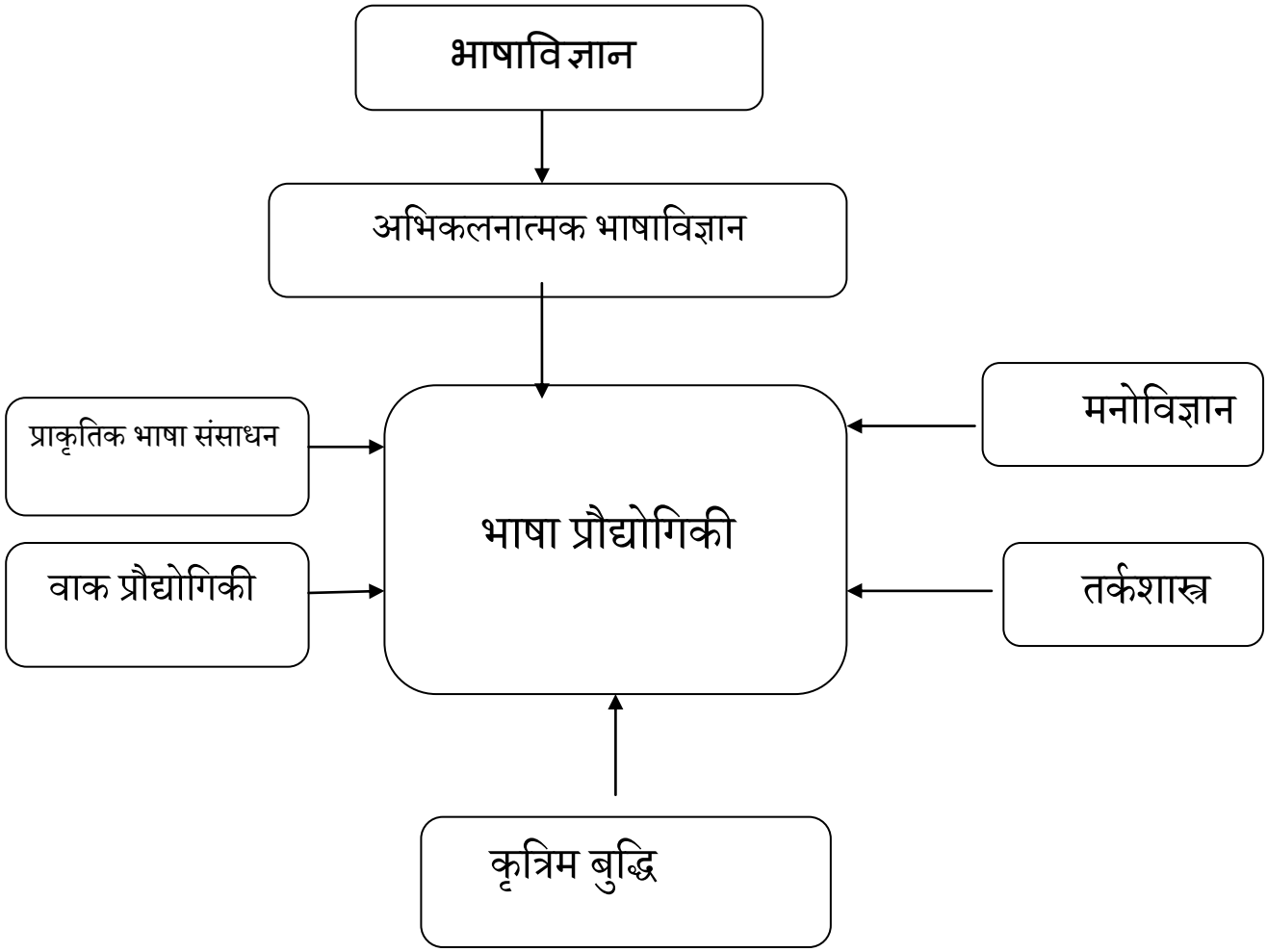
डॉ.कृष्णाकुमार गोस्वामी एक आलेख के अनुसार – भाषा प्रौद्योगिकी के मुल्तत्व अभिकलनात्मक भाषाविज्ञान और वाक प्रौद्योगिकी (speech technology) है किंतु इनमें कृत्रिम बुद्धि (A.I.) अर्थविज्ञान,गणित,तर्कशास्त्र आदि अन्य क्षेत्र भी परस्पर व्याप्त हैं।

"Language Technology, Wikipedia के अनुसार-

Language Technology is often called human language technology **(HLT)** or Natural Language Processing **(NLP)** and consists of Computational Linguistics and speech technology as its core but includes also many application oriented aspects of them. Language Technology connected to computer science and general linguistics."

इस प्रकार भाषा प्रौद्योगिकी के मूलतत्व अभिकलनात्मक भाषाविज्ञान तथा वाक प्रौद्योगिकी हैं किंतु इसमें कृत्रिम बुद्धि (A.I.), अर्थविज्ञान , गणित ,तर्क शस्त्र तथा दर्शन आदि अन्य क्षेत्र भी सहायता करते हैं। प्राकृतिक भाषा संसाधन इसका उपक्षेत्र है जिसमें प्राकृतिक मानव भाषाओं के स्वचालित जनन और बोधन से संबन्धित समस्याओं का अध्ययन होता है। प्राकृतिक भाषा जनन प्रणाली सूचना को कंप्यूटर डाटाबेस से सामान्य युक्तिपूर्ण मानव भाषा में परिवर्तित करती है और प्राकृतिक भाषा बोधन प्रणाली मानव के प्रतिदर्शों(samples) को अधिक औपचारिक प्रतिरूपों में परिवर्तित करती है जो कंप्यूटर प्रोग्रामों को सरलता से कम में लाती है। इस प्रकार भाषा प्रौद्योगिकी विभिन्न भाषा संसाधनों के विकास में सहायक है। इस का लक्ष्य सॉफ्टवेर उत्पादों का सृजन करना है। जिसमें मानव भाषा का कतिपय ज्ञान होता है।

भाषा प्रौद्योगिकी एक अंतरानुशासनिक विषय है , जिसमें भाषाविज्ञान गणित,दर्शन,कंप्यूटर,कृत्रिम बुद्धि,मनोविज्ञान जैसे विषयों का समावेश होता है। निम्न चित्र आरेख द्वारा इसे दिखा सकते हैं



कृत्रिम बुद्धि(Artificial Intelligence):-

कृत्रिम बुद्धि वस्तुतः मानव – बुद्धि की कार्यप्रणाली और निर्णय प्रक्रिया का कंप्यूटर अनुकरण है। ज्ञान तथा अनुभव जो मानव मस्तिष्क में अव्यक्त रूप में अंतर्निहित होते हैं, उनको व्यक्त रूप में कंप्यूटर की स्मृतिकोश में

डाला जाता है जिसके कारण कंप्यूटर वही निर्णय और आउपुट देता है जो मानव देता है।

सर्वप्रथम कृत्रिम बुद्धि की संकल्पना को **जॉन मेकार्थी** ने सन **1956** में प्रतिपादित किया था । तब से आज तक इस क्षेत्र में निरंतर कार्य हो रहे हैं। आज कृत्रिम बुद्धि के द्वारा मानव से भी द्रुत गति में कार्य करने में समर्थ है। परंतु किसी भी स्थिति में कृत्रिम बुद्धि मानव बुद्धि की स्थानापन्न नहीं हो सकती है।

अभिकलनात्मक भाषाविज्ञान के अंतगर्त प्राकृतिक भाषा संसाधन में कृत्रिम बुद्धि की महत्वपूर्ण भूमिका है। निरंतर बढ़ते तकनीकी के कारण मशीन में प्राकृतिक भाषा की समझ(Understanding), विश्लेषण(Analysis), जनन(Generation) और संप्रेषण(Communication) करने की क्षमता विकसित हो चुकी है। राष्ट्रीय और अंतरराष्ट्रीय स्तर पर कई कंप्यूटर वैज्ञानिक तथा भाषाविद कृत्रिम बुद्धि के आधार पर इस अवधारण के साथ ज्ञान – संसाधन (Knowledge Processing) एवं प्रबंधन(Management) के क्षेत्र में कार्य कर रहे हैं।

जिस प्रकार व्यक्ति किसी भाषा के समझने के लिए भाषिक घटकों जैसे- वाक , शब्द , पद , वाक्य , अर्थ और संदर्भ का तार्किक तथा नियमबद्धता के समन्वय से विचार व्यक्त करता है अथवा समझता है ठीक उसी प्रकार यदि कंप्यूटर को भी सिखाया जाय तो वह भाषा का विश्लेषण,जनन और संप्रेषण सिखाए गए ज्ञान के आधार पर कर सकता है। इस प्रकार भाषा प्रौद्योगिकी प्राकृतिक भाषा संसाधन की एक ऐसी विधा है जिसमें भाषाविज्ञान के अनुप्रयोगिक विषय अभिकलनात्मक

भाषाविज्ञान और वाक प्रौद्योगिकी का विशिष्ट योगदान है। इस के कारण सूचना प्रौद्योगिकी का क्षेत्र सुदृढ़ और सुगठित हो गया है। भाषा प्रौद्योगिकी के कारण ऐसे अनेक उपकरणों का विकास हो रहा है जिनकी भूमिका मानव कल्याण और राष्ट्रोत्थान के लिए महत्वपूर्ण है।

➢ शोध परिकल्पना :-

यह शोध कार्य मुख्य रूप से मशीनी अनुवाद पर आधारित होगा। इसके अंतगर्त मणिपुरी से हिंदी भाषा में parallel text दोनों भाषाओं का अध्ययन किया जायेगा। यह अध्ययन एक ही text को लेके काम करेगा । जो मणिपुरी text को हिंदी में अनुवाद करके दोनों भाषाओं कॉर्पस संरेखन(alignment) करेगें ।

➢ शोध का उद्देश्य :-

इस शोध कार्य का उद्देश्य मणिपुरी से हिंदी भाषाओं के मध्य परस्पर मशीनी अनुवाद को सुगम बनाना है। दोनों भाषाओं के मध्य अपसरण के अध्ययन से एक भाषा से दूसरी भाषा के बीच भिन्नताओं चिहिनत को करके उन के लिए नियम बनाए जा सकते हैं जो मणिपुरी , हिंदी दोनों भाषाओं के परस्पर आदान- प्रदान को बढ़ाएगा।

इसके अंतिरिक्त मणिपुरी से हिंदी व हिंदी से मणिपुरी भाषा सिखाने वालों के लिए यह अध्ययन उपयोगी साबित होगा।

➢ शोध प्रविधि :-

इस कार्य को करने के लिए उचित मात्रा में data collection के बाद निर्दशन पद्धति तथा कॉर्पस एवं data driven पद्धति का प्रयोग किया जायेगा। दोनों भाषाओं के वाक्यों का तुलनात्मक अध्ययन करके भिन्नताओं ज्ञात की जायेगी। यह शोध पूरी तरह से Experiment Based होगा।

➢ शोध का उपयोग:-

i. **Corpus as knowledge resource:** कॉर्पस का बहुभाषी लाइब्ररीस में, भाषा शिक्षण के लिए पाठ्य सामग्री बनाने में , एक भाषी शब्दकोश तैयार करने में ,बहुभाषी, द्विभाषी शब्दकोश तैयार करने में , मशीन पठनीय शब्दकोश बनाने में आदि उपयोग होता है।

ii. **Corpus in language Technology:** कॉर्पस का उपयोग वाक्य-विच्छेदन, आवृत्ति गणक ,टेक्स एनोटेशन, parts of speech tagging, information retrieval आदि के लिए tools design में होता है।

iii. **Corpus for human-machine interface system:** कॉर्पस का उपयोग OCR, Voice recognition, TTS, E-learning, on-line Teaching, question-answering इत्यादि के लिए होता है।

Iv. Corpus in speech Technology: स्पीच कॉर्पस का उपयोग ASR, speaker identification, forensic linguistics speech disorder में होता है।