Indian Streams Research Journal
ISSN 2230-7850
Volume-4 | Issue-4 | May-2014 Available online at www. isrj.net

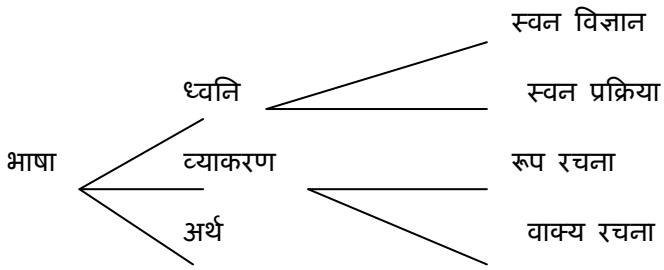
अनुक्रमिक पदच्छेदन अधिगम आधारित हिंदी नामीय पद अभिज्ञानक-'

अखिलेश कुमार
प्रौद्योगिकी अध्ययन केंद्र , मवर्धा.वि.हि.अं.गा.

सारांश :- भाषा मानव जीवन की अभिन्न अंग बन गई है। भाषा सर्वप्रथम संप्रेषण का कार्य निष्पन्न करती है। संगणकीय भाषाविज्ञान का उद्देश्य भाषा लिखित हो या मौखिक, इसकी अनंत संप्रेषण और सूचना संग्रह की क्षमता को संगणक द्वारा विश्लेषित करने में ही सन्निहित है। इसके लिए आवश्यक है कि संगणक को प्राकृतिक भाषा को समझने की क्षमता को विकसित किया जाए कि वह भाषा को सृजित करने व प्राकृतिक भाषा में निहित अनंत कार्यात्मक क्षमता को पहचानने में प्रशिक्षित हो सके। जब संगणक में यह क्षमता विकसित हो जाएगी , तब बहुत आसानी से संगणक द्वारा एक भाषा से दूसरी भाषा में भाषाई अनुवाद तथा प्राकृतिक भाषा की आवश्यक जानकारी हासिल करने में सफलता प्राप्त हो सकती है।

प्रस्तावना :

भाषा विज्ञान, जैसा कि इसके नाम से विदित होता है, भाषा का विज्ञान। इसके अंतर्गत भाषा एक अत्यंत जटिल संरचना है जिसके समस्त पक्षों व संपूर्ण संबंधों पर एक साथ दृष्टिपात करना एवं उनका सम्यक् वर्णन भाषाविद् के लिए संभव नहीं होता है। भाषा के अंतर्गत विभिन्न अध्ययन क्षेत्र को इस प्रकार देखा जा सकता है



उपरोक्त विभाजन में विद्वानों के अलग अलग मत हैं। संरचनात्मक व्याकरण के यूरोपीय विद्वान भाषा के-इन स्तरों को एक दूसरे से सुसंबद्ध कर भाषा को समग्र रूप में देखते हैं जबकि अमरीकी विद्वान इन स्तरों को एक दूसरे से अलग मानते हैं।

प्रस्तुत अध्ययन भाषाविज्ञान के सैद्धांतिक एवं संगणकीय पक्ष के विश्लेषण और वर्गीकरण को दर्शाता है। वर्तमान समय में भाषा विज्ञान को संगणक से संबद्ध कर नए विषय क्षेत्रों का उद्भव हुआ जिनमें भाषा - प्रौद्योगिकी, कंप्यूटेशनल भाषा विज्ञान, भाषा अभियांत्रिकी प्रमुख हैं। इन विषय क्षेत्रों में भाषा विज्ञान के नियमों एवं व्याकरण

को संगणक के संदर्भ में तैयार कर उसे संगणक द्वारा संसाधित कर मानवीय पक्ष को आत्मसात करने का प्रयास है। भाषा विज्ञान तथा हिंदी भाषा के विविध पक्षों का अध्ययन करते हुए जब अनुप्रयुक्त भाषा विज्ञान (Applied Linguistics पर दृष्टि जाती है तो अनुवाद का अध्ययन इसका एक अत्यंत स्वाभाविक क्षेत्र लगता है। अनुवाद (ध्वनि- प्रक्रिया का माध्यम भाषा होती है। भाषा के पक्ष से भाषा के विभिन्न स्तरों, शब्द, रूप, वाक्य, अर्थ, शैली आदि का वैज्ञानिक अध्ययन अुनवाद प्रक्रिया में किस प्रकार और कहाँ तक सहायक सिद्ध होते हैं , इन सब विषयों के संगणकीय पक्ष में किस प्रकार नियमों और प्रणाली विकास के लिए एल्गोरिद्म के द्वारा शामिल किया जाता है। इन सब विषयों का संक्षिप्त वर्णन प्रस्तुत अध्ययन में देखा गया है और प्रणाली विकास किया गया है।

प्राकृतिक भाषा संसाधन)NLP बहुत ही महत्वपूर्ण एवं चुनौती भरे विषय के रूप (में उभरा है। प्राकृति भाषा संसाधन के अंतर्गत निम्नलिखित क्षेत्रों को देखा जा सकता है शब्द कोटि निर्धारक)Pos Tagger(, शब्द विश्लेषक)Morphological Analyzer(, पद-विच्छेदन)Passer(, शब्द आवृत्तिगणक)Word frequency counter(, शब्द समूह)Chunker(, मशीन अनुवाद)Machine Translation(, Name Entity Recognizer नामीय पद अभिज्ञानक आदि।

मानव जाति को अपनी बात एक दूसरे के पास पहुँचाने के लिए भाषा का सहारा लेना पड़ता है। भाषा- प्रथमत संप्रेषण का कार्य निष्पन्न करती है। भाषा एक ऐसा माध्यम है जिससे संप्रेषण :, लिखित हो या मौखिक; सबसे कारगर सिद्ध होती है। संप्रेषण के दौरान अगर दूसरे देश या क्षेत्र में हो रही घटना को , उस देश या क्षेत्र की भाषा में समझ पाना मुश्किल हो जाता है तो परिस्थतियों को समझने और उपयोग में लाने के लिए अनुवाद की आवश्यकता महसूस होती है। विदेशी भाषा में मुद्रित सामग्री को समझने या किसी विदेशी ज्ञान - विज्ञान को उपयोग में लाने के लिए, अनुवादक की आवश्यकता होती है। परा आधुनिक युग वैश्वीकरण का युग- है। विश्व के सभी लोग मिलकर ज्ञान प्रमुख 70 विज्ञान के विकास के लिए प्रयत्नशील हैं। विश्व में लगभग- विज्ञान के विषय उपलब्ध हैं। एक भाषा से दूसरी भाषा में सामग्री को-भाषाएँ ऐसी हैं जिनमें साहित्य तथा ज्ञान ले जाने के लिए कुशल अनुवादक की आवश्यकता होती है। अर्थात अनुवाद के अभाव में आज विश्व के सचेत मानव अचेत हो सकते हैं। ज्ञान विज्ञान के अनेक लाभों से वंचित रह सकते हैं।

भाषाविज्ञान तथा हिंदी भाषा के विविध पक्षों का अध्ययन करते हुए जब अनुप्रयुक्त भाषाविज्ञान)Applied Linguisticsपर दृष्टि जाती है तो अनुवाद का अध्ययन इसका एक अत्यंत स्वाभावि (क अंग लगता है। अनुवाद की प्रक्रिया का माध्यम भाषा होती है। भाषा के पक्ष से भाषा के विभिन्न स्तरों ध्वनि -, शब्द, रूप, वाक्य, अर्थ, शैली आदि का वैज्ञानिक अध्ययन , अनुवाद प्रक्रिया में किस प्रकार और कहाँ तक सहायक सिद्ध होते हैं, यह प्रस्तुत शोधप्रबंध का केंद्-र है।

प्रस्तुत शोध प्रबंध का विषय-*'अनुक्रमिक पद्विच्छेदन अधिगम आधारित हिंदी नामीय पद अभिज्ञानक-'* **(Sequential Parsing Approach to Hindi Name Entity Recognizer)** है। अतः इस विषय पर किया गया शोध कार्य कहीं न कहीं मशीन अनुवाद से संबंध रखता है। शोध विषय में अनुक्रमिक से तात्पर्य अनुक्रमिक व्याकरण से है

जो कि किसी भाषा विशेष के लिए नहीं बल्कि इसके आधार पर किसी भी भाषा का पद विच्छेदन करने हेतु किया-) जा सकता है क्योंकि यह व्यापक उपागम General Approach पर आधारित है। (

नामीय पद अभिज्ञानक से तात्पर्य है कि किसी भी हिंदी पाठ के वाक्य संरचना में आए हुए नामीय पद को पहचान करना। किसी वाक्य में नामीय पद को पहचान मानव अपने ज्ञान और संदर्भ के आधार पर पहचान कर लेता है, परंतु मशीन के लिए एक सामान्य शब्द की तरह है और बिना नियमों या संदर्भ के मशीन के लिए एक चुनौती पूर्ण कार्य है। चूँकी संगणक पूजा ने पूजा के लिए फूल तोड़े उक्त वाक्य में निर्णय लेने में असफल हो जाता है कि किसे नाम समझे और किसे कार्य। इसी वाक्य को अगर पूजा के लिए पूजा ने फूल तोड़े इनपुट के रूप में दिया जाए तो बिना word sense disambiguate के नहीं समझ सकता। इसी प्रकार की समस्याओं को प्रस्तुत शोध के माध्यम से सुलझाने का एक प्रयत्न है।

उपयुक्त विषय पर इंग्लैंड कंप्यूटेशनल शोध समूह यू में अंग्रेजी भाषा एवं भारतीय परिपेक्ष्य में अन्य .के. भारतीय भाषाओं पर शोध कार्य हुए हैं एवं किए जा रहे हैं।

नामीय पद अभिज्ञानक विषय मात्र एक या दो दशक पुराना है। प्राकृतिक भाषा संसाधन में आ रही कठिनाईयों पर कार्य करते हुए भाषा वैज्ञानिकों का ध्यान नामीय पद अभिज्ञानक की ओर गया। सन 1965 में MUC-6 में उपस्थित प्रतिभागियों के समक्ष इसकी परिकल्पना रखी गयी। इसके लिए उस समय अंग्रेजी भाषा का चुनाव किया गया। अंग्रेजी भाषा को ध्यान में रखते हुए इसके लिए तकनीक विकसित की गई, जो काफी हद तक इस तरह की समस्याओं के समाधान में सक्षम था।

यह तकनीक हिंदी में अपने प्रारंभिक दौर में है। इस विषय पर अब तक हुए शोध कार्यों का अवलोकन करने पर ज्ञात होता है कि निम्नलिखित संस्थान इस विषय पर कार्यरत हैं। सुदेषना सरकार, सूजन कुमार साहा, पार्थ सारथी घोष और पवित्र मि त्रा ने मिलकर हिंदी के लिए पहला नामीय अभिज्ञानक बनाया जो Maximum entropy and Translation सिद्धांत पर आधारित था। कुछ प्रमुख संस्थान भारतीय भाषाओं के लिए नामीय अभिज्ञानक को विकसित करने के लिए कार्यरत हैं। उनमें कुछ प्रमुख संस्थान निम्नलिखित हैं

- Language technology Lab, IIIT Hyderabad
- IIT Bombay
- Center for Sanskrit Computational Linguistics, JNU, New Delhi
- CIIL, Mysore
- MGAHV, Wardha
- TDIL
- C-DAC Noida
- HCV, Hyderabad
- Jadhavpur University

प्रस्तुत शोधप्रबंध अनुक्रमिक अधिगम पर आधारित है।-

1.1 प्रथमत :हिंदी व्याकरण का अध्ययन किया गया है जिसमें हिंदी व्याकरण की पृष्ठभूमि , विकास क्रम, इतिहास, अन्य भाषाओं के साथ संबंध , व्याकरण निर्माण की प्रक्रिया आदि पर समग्र अध्ययन प्रस्तुत करने का प्रयास किया गया है। द्वितीय उपशीर्षक हिंदी भाषा की प्रकृति के अंतर्गत हिंदी की कुछ प्रकृतिगत विशेषताओं को चिह्नित किया गया है। तृतीय उपशीर्षक हिंदी का शब्द भंडार के- अंतर्गत हिंदी ने कुछ विदेशी भाषाओं के भी शब्द लिए हैं। सभी चलायमान भाषाएं ऐसा करती हैंनजदीक आयी किसी भी विदेशी भाषा के आवश्यक शब्द ग्रहण कर - लेती हैं। किंतु कोई भी भाषा अपने क्रियापद, सर्वनाम तथा विभक्तियाँ आदि नहीं बदलती। चतुर्थ उपशीर्षक शब्द और शब्दभेद के अंतर्गत शब्द निर्माण, पद निर्माण, तत्सम, तद्भव, देशज, विदेशी शब्द का अध्ययन किया गया है। पंचम उपशीर्षक में हिंदी शब्दभेद का विस्तृत अध्ययन करने का प्रयास है। अंत में नामीय पद विश्लेषण की परिभाषा, स्वरूप की चर्चा की गई है।

1.2 नामीय पदअभिज्ञा :विच्छेदन-न प्रक्रिया को तीन उपशीर्षकों के अंतर्गत निम्नलिखित बिंदुओं का अध्ययन किया गया है। पद विच्छेदन प्रक्रिया-, विविध अधिगम, अनुक्रमिक पद विच्छेदन- विच्छेदन आदि। नामीय पद- अभिज्ञान प्रक्रिया'हिंदी)NEER)' से तात्पर्य मुख्यतः हिंदी वाक्यों में आए नामीय पदों का अभिज्ञान पद विच्छेदन- प्रक्रिया द्वारा करने से है।

पद विच्छेदन अभिज्ञान एक ऐसी प्रक्रिया है जिसमें वाक्य संरचना के अंतर्गत आयी सभी व्याकरणिक-) कोटियों का अंतःसंबंध के आधार पर विच्छेदनPar si ngकरना होता है। (

प्राकृतिक भाषा का वर्णन करने के लिए कई व्याकरण सिद्धांतों का विकास किया गया है और स्वतः पद - विच्छेदन के कुछ प्रकार भाषा के संगणकीय विश्लेषण में काफी प्रभावशाली रहे हैं। व्याकरण के सिद्धांतों में से कुछ संदर्भों के लिए इन सिद्धांतों की समीक्षा की गई है।

(क (पदबंध संरचना व्याकरण)Phr ase st r uct ur e gr ammar (

(ख (वृक्ष संरचना व्याकरण)Tr ee adj oi ni g gr ammar (

(ग (कोटिगत व्याकरण)Cat egor i cal gr ammar (

(घ (निर्भरता व्याकरण)dependency gr ammar (

(ङ (रूपांतरण व्याकरण)Tr ansf or mat i onal gr ammar (

(च (नियम और बंधन सिद्धांत)Gover nment and bi ndi ng t heor y(

व्याकरण का उदाहरण

पदविच्छेदन के लिए व्याकरण का उदारहण-

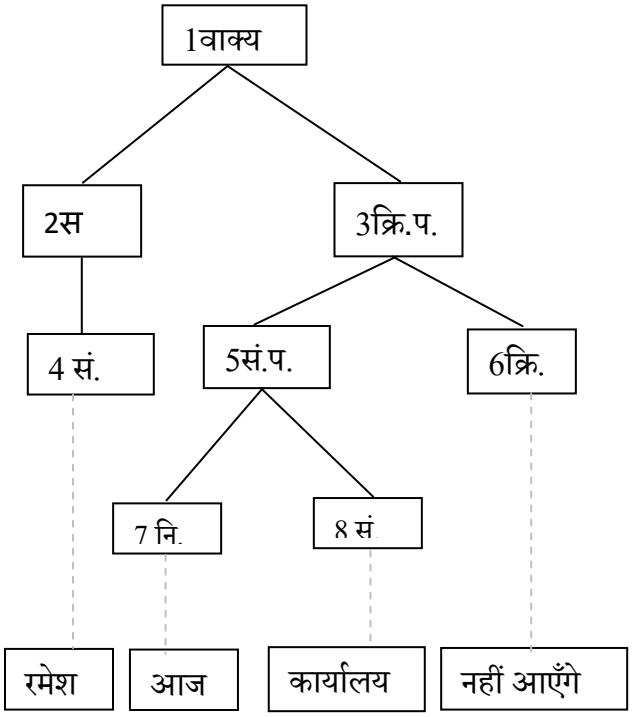
S - -> NP VP S = Sent ence (वाक्य(
NP - -> DET N NP = Noun Phr ase (संज्ञा पदबंध(
VP - -> V N P DET- Det er mi ner (निर्धारक(
VP - -> V NP N = Noun (संज्ञा(
VP = Ver b Phr ase (क्रिया पदबंध(
V = Ver b (क्रिया((क (व्याकरणिक नियम
NP - -> वह
NP - -> रमेश
V - -> देखा
V - -> है
N - -> बच्चाम
N - -> लड़का
DET-->एक

आरेख-1 एवं 2

पदबंध संरचना नियम
वाक्य > सं.प.+क्रि.प.
सं .प.> सं.
सं .प.> नि.+सं.
क्रि .प.> सं.प.+क्रि

1. वृक्ष संरचना

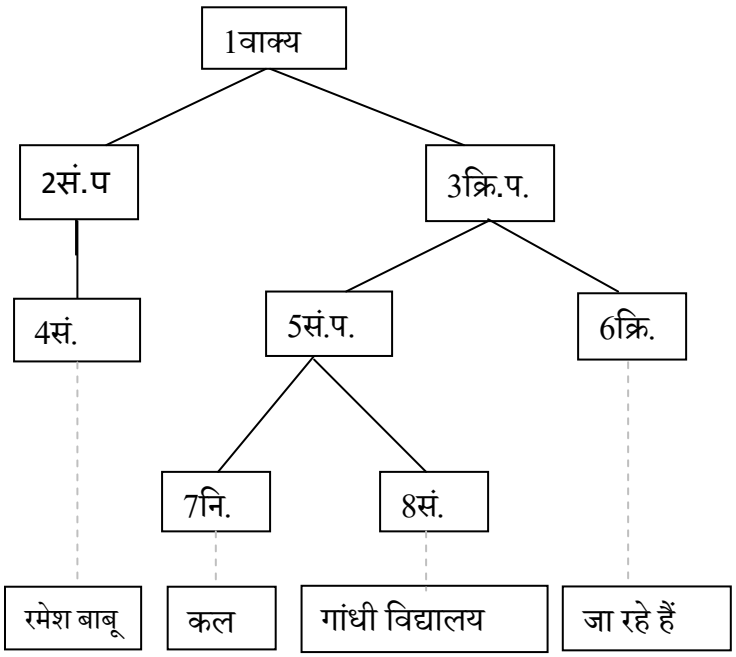
पदबंध संरचना नियम
वाक्य.सं-प.+क्रि.प.
सं.प.सं-.
सं.प.नि-.+सं.
क्रि.प.सं-.प.+क्रि.
शब्दकोश (Lexicon)
सं(रमेश) .
नि(आज) .
सं(कार्यालय) .
क्रि(नहीं आएंगे) .



2. वृक्ष संरचना

3. वृक्ष संरचना

पदबंध संरचना नियम
वाक्य.सं-प.+क्रि.
सं.प.सं-.
सं.प.नि-.+सं.
क्रि.प.सं-.प.+क्रि.
शब्दकोश (Lexicon)
सं(रमेश बाबू) .
नि(कल) .
सं(गांधी विद्यालय) .
क्रि((जा रहे हैं) .



दिए गए व्युत्पन्न वृक्ष संरचना में अनुक्रमिक पद विच्छेदन को दर्शाने का-प्रयास किया गया है।

पद) विच्छेदन के विविध अधिगमों के अंतर्गत पदबंध संरचना व्याकरण-Phrase Structure Grammar(, वृक्ष संरचना व्याकरण)Tree adjoining Grammar(, कोटिगत व्याकरण)Categorical Grammar(, निर्भरता व्याकरण)Dependency Grammar(, रूपांतरण व्याकरण)Transformational Grammar(, नियम और बंधन सिद्धांत)Government and binding theoryआदि की चर्चा की गई है। (

अन्य उपशीर्षकों सांखिकीय-, भाषा वैज्ञानिक, नियमाधारित, कोश आधारित, संकर आदि को परिभाषित करते हुए प्रस्तुत शोध कार्य के अंतर्गत प्रणाली विकास में प्रयोग किए गए नियमों के निर्माण संबंधी जानकारी प्रस्तुत की गई है। उपशीर्षक शब्द और उसके विश्लेषण के अंतर्गत प्राकृतिक भाषा संसाधन में प्रजनक और - विश्लेषक किस प्रकार कार्य निष्पादन करते हैं, देखने का प्रयास किया गया है।

रूपिमीय विश्लेषण को क्रमशः कुछ एल्गोरिद्म और चित्र की सहायता से प्रस्तुत करने का प्रयास किया गया है।

अंत में हिंदी संज्ञा पदबंध के घटक, हिंदी में निर्धारक, हिंदी में पूर्व एवं पश्च विशेषक की चर्चा की गई है।

1.3 'डाटाबेस निर्माण एवं प्रबंधन' के अंतर्गत इससे संबंधित परिचय, डाटाबेस संरचना तथा उपयोगिता, 'डाटाबेस मैनेजमेंट सिस्टम, डाटाबेस संरचना, डाटाबेस के तत्व, डाटाबेस डिजाइन एवं प्रस्तुतीकरण, डाटाबेस मॉडल, स्ट्रक्चर्ड क्वेरी लैंग्वेज, डाटा परिचालन, ट्रांजेक्शन कंट्रोल, डाटा निरूपण, आदि को दर्शाया गया है।

डाटाबेस संकलित किए गए आँकड़ों का एक समूह है। इसी प्रकार सभी उपशीर्षकों को परिभाषित करने का प्रयास चित्रों एवं कोडों के माध्यम से किया गया है।

प्रस्तुत शोध कार्य के अंतर्गत विकसित प्रणाली नियमों के आधार पर कार्य करती है परंतु इसमें SQL Server 2005 का उपयोग किया गया है जो कि प्रत्येक वाक्य को पूर्ण विराम के आधार पर पहचान करता है। इसके बाद प्रत्येक वाक्य के शब्द को इस टेबल में संग्रहीत करता है तथा नियमों के आधार पर उस शब्द का वर्णन और कोटि आदि का निर्धारण करता है। इसी प्रकार प्रणाली प्रत्येक वाक्य के साथ कार्य करता है।

डाटाबेस डिजाइन एवं प्रस्तुतीकरण के अंतर्गत)SQL Server 2005) प्रारंभ से लेकर विकसित प्रणाली में डाटा संग्रह एवं कार्य प्रणाली को चरणबद्ध तरीके से प्रस्तुत करने का प्रयास किया गया है। डाटाबेस मॉडल के अंतर्गत मॉडल के विभिन्न प्रकार एवं प्रणाली किस मॉडल पर कार्य करती है प्रस्तुत किया गया है। संरचित क्वेरी भाषा की चर्चा करते हुए भाषा के तत्व को प्रस्तुत किया गया है।

अंत में डाटा परिचालन, ट्रांजेक्शन कंट्रोल्स, डाटा निरूपण, डाटा कंट्रोल आदि को प्रणाली में कार्य कर रहे कोडों के द्वारा प्रस्तुत करने का प्रयास किया गया है कि प्रणाली डाटा को किन किन चरणों से ह-ोकर गुजरती है और किस प्रकार उसे निर्मित किया जाता है। उपरोक्त सभी का Syntax दिया गया है।

1.4 ''हिंदी)NER)' संसाधन एवं प्रकार्य का परिचय :, कार्य प्रणाली, प्रणाली संरचना, कलनविधि-, प्रवाह सचित्र, 'हिंदी)NER)' विकास एवं प्रकार्य, प्रस्तुतीकरण, अंतरापृष्ठ विकास, चरणबद्ध प्रचालन आदि विषयों को क्रमबद्ध रूप से अध्ययन किया गया है।

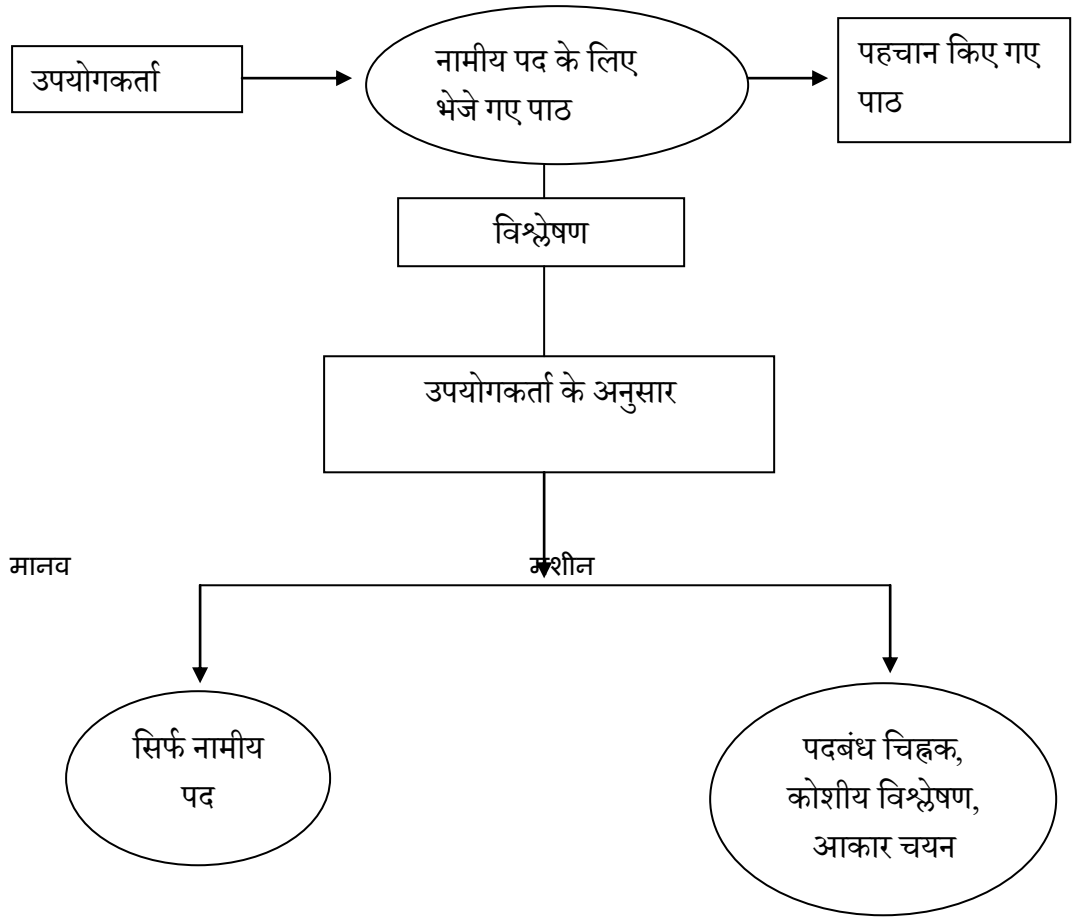
'हिंदी)NER)' प्रणाली हिंदी वाक्यों में आए नामीय पदों की पहचान कर उसे चिह्नित करता है। सामान्यतः प्रणाली विकास किसी उद्देश्य की पूर्ति हेतु किया जाता है। यहाँ नामीय पदों के विश्लेषण एवं पहचान का मुख्य

उद्देश्य संगणक के द्वारा एक ऐसे उपकरण का विकास करने से है जो मानव की तरह हिंदी वाक्यों में आए नामीय पदों की पहचान कर सके और भविष्य में मशीन अनुवाद जैसे भावी लक्ष्य को पूरा करने में सहायता प्रदान कर सके। 'हिंदी)NER)' के कार्य प्रणाली वेब डिजाइन , Visual Studio 2008 ASP.Net और SQL Server 2005 के सम्मिलित रूप से तैयार किया गया एक प्रोग्राम है, जो पूर्णतः नियमों पर निर्भर करता है। परंतु कुछ अपवादों की समस्या को लेकर आंशिक रूप से डाटाबेस पर निर्भर करती है।

प्रणाली को सर्वप्रथम हिंदी वाक्यों का पाठ, पठन के लिए दिया जाता है जो कि पठन के बाद पदबंध चिह्नित करता है। फिर उसे टोकनीकृत कर कोटि पहचान करता है , पुनः कोटि लेवलिंग कर नामीय पदों की पहचान कर कोटिकृत शब्दों को आउटपुट के रूप में देता है।

1. प्रणाली संरचना विशिष्टता

डाटा फ्लो प्रारूप



कलन विधि) -:Algorithm)

किसी भी प्रणाली के एल्गोरिद्म से तात्पर्य प्रणाली की समस्या के समाधान हेतु विकास क्रम में लिए गए सूचनाओं के तार्किक एवं व्यवस्थित रूप से है। प्रणाली विकास के दौरान इच्छित परिणाम प्राप्त करने हेतु इसे छोटे छोटे वाक्यों के व्यवस्थित क्रम में लिखा जाता है। प्रणाली का परिणाम मुख्य रूप से एल्गोरिद्म पर निर्भर करता है। निम्नलिखित निर्देश किसी भी प्रणाली विकास के लिए महत्वपूर्ण होते हैं -

- निर्देश छोटा और संपूर्ण होना चाहिए।
- निर्देश दिए गए समय में परिणाम देता हो।
- एल्गोरिद्म का निर्देश निश्चित स्थान पर समाप्त होता हो।
- निर्देशों का अंतहीन संरचना में दुहराव नहीं होता हो।
- परिणाम प्राप्त होने के पश्चात एल्गोरिद्म समाप्त होना चाहिए।

(1) आगत वाक्य का पठन)To Read Input Sentences)।

(2) पठित वाक्य का पदबंध चिन्हन नाम तत्व प्रत्यभिज्ञान।/

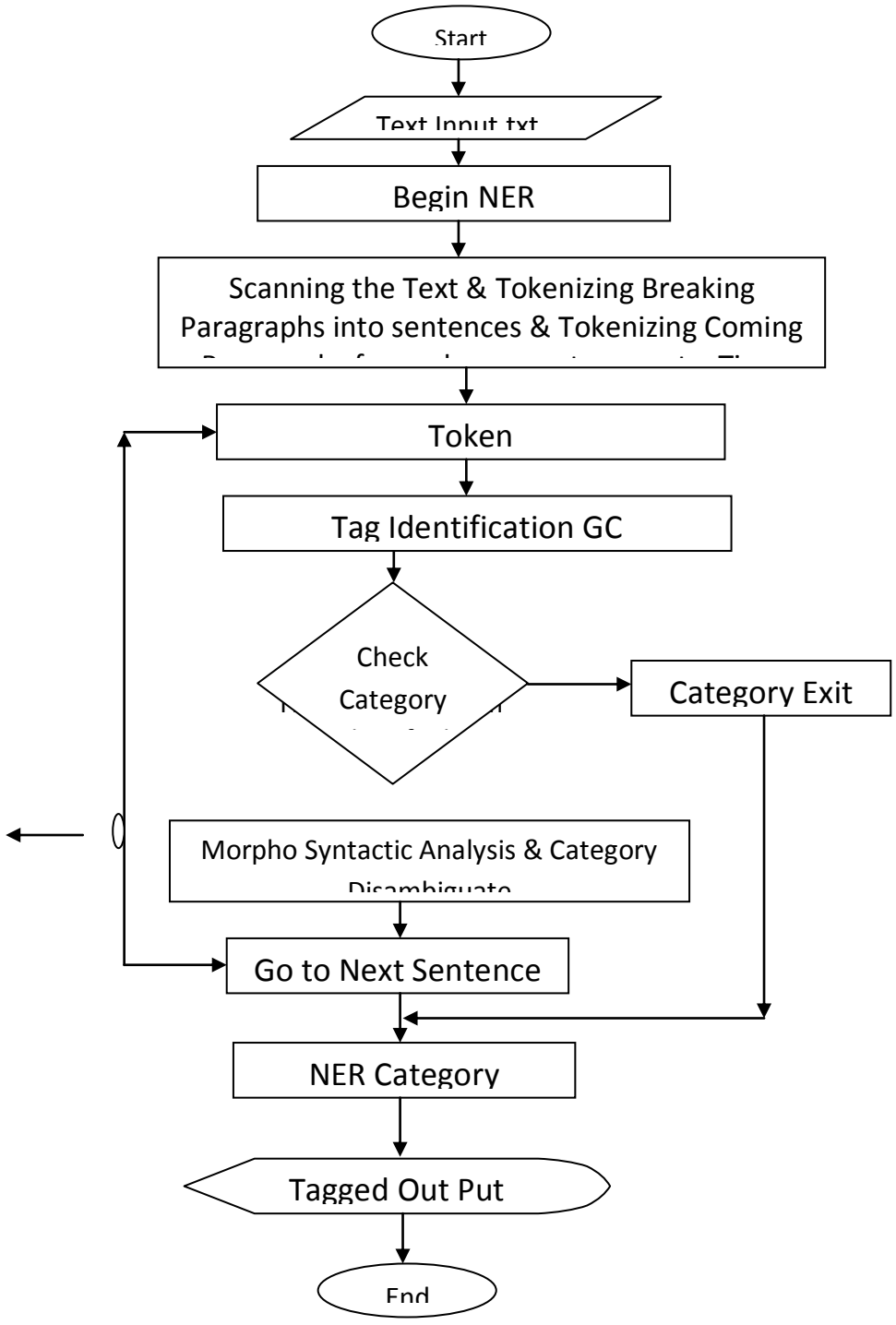
(To Phrase - Marking Name Entity Recognizing)।

(3) टैग निर्धारित करना)Assign The Tags)।

- पूर्व में लिए गए शब्दवृत डेटाबेस द्वारा कोटि अभिज्ञात होना और दिए गए नियमों का अनुपालन- अभिज्ञान करती है तो कोटि- करना। यदि नियम केवल एक कोटि के रूप में शब्द का कोटि उपलब्धता, अन्यथा वाक्य में एकल शब्द की आवृति द्वारा दो कोटियों के होने की (समनाम) संभावना और अगले चरण में प्रवेश।
- आए हुए वाक्य पर संभव वाक्यविन्यासी नियम लागू करना और नामीय पद कोटि प्राप्त होना।-
- प्राप्त कोटि का पुर्नअंकन करना)To Label Fined Category)।

प्रस्तुत शोध कार्य में प्रणाली विकास के दौरान नियमों का निर्माण अधिक जटिल समस्या थी। इन नियमों के निर्माण में जहाँ एक नियम सफल होती है वहीं दूसरे हिंदी वाक्य संरचना में असफल हो जाती है। इसी प्रकार की अनेक समस्याओं को सुलझा कर प्रणाली विकास करना सबसे अधिक समय लगा। चूँकि नियमों पर सफलता ही प्रणाली की सफलता को प्रमाणित करती है।

प्रवाह सचित्र (**Flow Chart**)



'हिंदी)NER)' विकास एवं प्रकार्य में विकसित प्रणाली के उपयोग में लिए गए सॉफ्टवेयर की चर्चा एवं प्रणाली संचालन को क्रमबद्ध तरीके से प्रस्तुत किया गया है।

प्रस्तुतीकरण के अंतर्गत प्रणाली के इंटरफेस का स्क्रीन शॉट दिखाया गया है , जिसमें इनपुट वाक्य और आउटपुट वाक्यों में 'हिंदी)NER)' को पहचान करते हुए दिखाया गया है। चरणबद्ध प्रचालन के अंतर्गत विकसित प्रणाली के प्रत्येक इंटरफेस को क्रमबद्ध रूप से दिखाया गया है जिसे विकसित प्रणाली के संचालन के अनुरूप स्क्रीन शॉट के माध्यम से प्रस्तुत किया गया है।

1.5 इस शोध का कार्य मूल्यांकन के अंतर्गत प्रतिदर्श संकलन, प्रकृति, संकलन मानदंड, प्रतिदर्श, आकार, मूल्यांकन विधि, प्रणाली प्रस्तुतीकरण , परिशुद्धता, समयनिमित-, जटिलता निवारण , क्षेत्रानुसार सांख्यिकीय प्रस्तुतीकरण उपशीर्षकों के अंतर्गत विश्लेषण करने का प्रयास किया गया है।

प्रतिदर्श संकलन- हिंदी उपन्यासों एवं कुछ जटिल वाक्यों के निर्माण द्वारा सफलता हासिल करने का प्रयास किया गया है। प्रतिदर्श की प्रकृति नियमों के निर्माण हेतु हिंदी वाक्यों में नाम से चिह्नित वाक्यों को किताबों एवं- कुछ अन्य संस्थानों में जिन वाक्यों को लेकर नियम निर्माण की प्रक्रिया जारी है उन वाक्यों को प्राथमिकता देते हुए नियमों का निर्माण किया गया है।

संकलन मानदंड- हिंदी साहित्य को चुनने का खास मकसद यह था कि प्रणाली को दिए गए पाठ में अधिकअधिक नाम मिले। अतः कर्मभूमि उपन्यास को लिया गया जिसमें पात्रों की संख्या अधि-से-क है तथा सबके अलग अलग नाम हैं।-

प्रणाली आकार- प्रशिक्षित डाटा के लिए 200 वाक्यों का निर्माण किया गया जिसके आधार पर प्रणाली को प्रशिक्षित किया गया एवं कर्मभूमि)144225) शब्द का डाटा लिया गया।

मूल्यांकन विधि- प्रस्तुत ''हिंदी)NER)' ' प्रणाली का मूल्यांकन 12 व्यक्तियों से प्रणाली परीक्षण कराया गया।

प्रणाली प्रस्तुतीकरण-- प्रस्तुत प्रणाली 'हिंदी)NER)' हिंदी वाक्यों में आए नामीय पदों को दो स्तर पर प्रदर्शित करती है। जटिलता निवारण व्याकरणिक कोटियों के आधार पर नियमों को संगणकीय संदर्भ में लागू - करते समय कुछ अपवाद भी प्रस्तुत हुए। इन समस्याओं के समाधान हेतु इन अपवादों को नियमों द्वारा समाप्त करने का प्रयास किया गया है।

क्षेत्रानुसार सांख्किीय प्रस्तुतीकरण के अंतर्गत मूल्यांकन पत्र के आधार पर प्राप्त आंकड़ों को ग्राफ चित्र द्वारा प्रस्तुत करने का प्रयास किया गया है।

1. 6प्राप्ति

प्रस्तुत NER प्रणाली दिए गए हिंदी वाक्यों में नामीय पद की पहचान सफलता पूर्वक करता है। प्रस्तुत शोध कार्य हिंदी भाषा के लिए है। वर्तमान समय में मशीन अनुवाद और कंप्यूटेशनल भाषाविज्ञान के क्षेत्र में आ रहे समस्याओं के लिए महत्वपूर्ण साबित हो सकती है।

1. उपयोगिता 7

प्रणाली प्राकृतिक भाषा संसाधन के कई प्रमुख क्षेत्रों में मदद्गार साबित होगा जिसमें सर्व प्रथम यह प्रणाली अपने विषय नामीय पद अभिज्ञानक के अलावा टैगर, वर्तनी जाँचक, सूचना पुनर्प्राप्ति, प्रोक्ति विश्लेषण, पाठ विश्लेषण, ड्रामा, उपन्यास, हिंदी और विदेशी भाषा शिक्षण)FLT(आदि क्षेत्रों में सहायक सिद्ध होगा।

1.8. 1सैद्धांतिक

पदविच्छेदन प-ुरक्रिया कोश आधारित से तात्पर्य यह है कि नामीय पदों का स मस्या डाटाबेस में बनाना और फिर प्रोग्रामिंग के द्वारा उसे मिलान करते हुए अभिज्ञान कराना। पदविच्छेदन प्रक्रिया कोश आधारित होने पर- गलत होने की संभावना शून्य होती है परंतु नामीय पदों की संख्या अनगिनत है और जब हम इसका डाटाबेस तैयार करेंगे तो यह काफी बड़ा हो जाएगा और किसी भी शब्द के अभिज्ञान के लिए प्रणाली को पूरे डाटाबेस से मिलान करना होगा जिसमें काफी समय लगने के कारण यह प्रणाली परिणाम देने में बहुत समय लेगा जो कि आज के आधुनिक युग के लिए बेकार साबित होगा। दूसरी बात इस प्रकार का डाटाबेस बनाने में भी सभी डोमेन और द्विअर्थकता या संदिग्र्थाथकता पर सटीक नहीं बैठेगा और असफल हो जाएगा।

1.8.2. अनुप्रयोगिक

पद विच्छेदन प्रक्रिया में नामीय पदों को पहचानने के लिए नियमों का निर्माण किया गया है। संगणक- इन्हीं नियमों के आधार पर पद विच्छेदन प्रक्रिया के बाद नामीय पदों का अभिज्ञान करने में संभव हो पाता है।- प्रणाली इन्हीं नियमों के आधार पर कार्य करता है परंतु इन नियमों से अलग ज्यों ही किसी वाक्य में अंतर होता है तो प्रणाली असफल हो जाता है। अतः नियमों को इतना तार्किक ढंग से सोचा गया है कि नियम असफल न होने पाए। जहाँ इस तरह की समस्या दिखाई दे तो प्रणाली को अपवाद या प्रोग्रामिंग में एरे के द्वारा उन शब्दों को जोड़कर उन समस्याओं से निजात पाई जाती है। उदाहरण स्वरूप कुछ नियम निम्नलिखित हैं।

सम्मान सूचक शब्दों, जैसे.डॉ -, प्रो., मास्टर, मि., श्री, श्रीमती, श्रीमान, महोदय, महाशय, माननीय, सुश्री आदि के बाद आनेवाले शब्द निश्चित तौर पर नाम होंगे लेकिन कुछ सम्मानित शब्द जैसे महोदय -, महाशय के बाद कभीकभी नाम नहीं हो सकता-, जैसेकिसी आवेदन पत्र आदि में। - प्रस्तुत शोध में इस प्रकार कई नियमों को देकर प्रणाली विकास किया गया है। प्रस्तुत प्रणाली हिंदी वाक्य में आए नामीय पदों को सफलता पूर्वक टैग करता है।

1. 9सीमा

प्रस्तुत प्रणाली अभी ट्रेंड डाटा पर चलाया गया है जो साहित्य के दो पुस्तकों पर पूर्णरूपेण 95% तक शुद्धता प्रदर्शित करती है। अत यह अनुमान :किया जा सकता है कि उन्हीं लेखकों के अन्य पुस्तकों पर ये शुद्धता विद्यमान रहेगी अथवा शुद्धता की प्रतिशत कम हो सकती है। इस प्रणाली को ट्रेंड करने के लिए जटिल वाक्यों का चयन किया गया है। परंतु हिंदी वाक्यों की संरचना बदलते ही अर्थ में परिवर्तन संभव है। अत इस :की कुल सीमा ट्रेंड डाटा के अलावा 95 पृष्ठों पर चेक किया गया जिसमें इस प्रणाली की शुद्धता 900% रही।

1. 10संभावना

प्रस्तुत प्रणाली को भविष्य में प्रत्येक प्रकार के डाटा द्वारा प्रशिक्षित कराकर 100% शुद्धता प्राप्त करते हुए इसे मशीन अनुवाद प्रणाली के लिए उपयोग किया जा सकता है। आनेवाले समय में शोधार्थी इस विषय को पूर्णत: नियमों पर आधारित कर प्रणाली विकास कार्य को आगे बढ़ा सकते हैं।

हिंदी संदर्भ सूची-ग्रंथ-

- ओझा, त्रिभुवन)1986) : हिंदी में अनेकार्थता का अनुशीलन , विश्वविद्यालय प्रकाशन , वाराणसी , भारत।
- गुरू, कामता प्रसाद)1920) : हिंदी व्याकरण, काशी नागरी प्रचारिणी सभा, वाराणसी, भारत।
- जैन, वृषभ प्रसाद)1995) : अनुवाद और मशीनी अनुवाद, सांराश प्रकाशन ,नई दिल्ली, भारत।
- द्विवेदी) कपिल देव ,2002) भाषाविज्ञान एवं भाषा शास्त्र, विश्वविद्यालय प्रकाशन, वाराणसी, भारत।
- पांडेय, प्रो) महेंद्र कुमार .2008-अंक) बहुवचन : (18प्रकाशन .वि.हिं.अं.गां.म (, भारत।
- वाजपेयी, आचार्य किशोरी दास)1985 नागरी प्रचारिणी सभा :हिंदी शब्दानुशासन : (, वाराणसी, भारत।
- शर्मा, डॉ) रामकिशोर .2004) : आधुनिक भाषाविज्ञान के सिद्धान्त (पंचम संस्करण), इलाहाबाद , लोकभारती प्रकाशन।
- सिंह) सूरजभान ,2000) : हिंदी का वाक्यात्मक व्याकरण , दिल्ली, साहित्य सहकार, नयी दिल्ली , भारत।
- सिंह) सूरजभान ,2003) : अंग्रेजीहिंदी अनुवाद व्याकरण-, प्रभात प्रकाशन, नयी दिल्ली, भारत।
- सिन्हा, लक्ष्मण प्रसाद)198 3हिंदी भाषा का रूपिमीय विश्लेषण : (, अंशुकमल प्रकाशन, पटना, भारत।

अंग्रेजी संदर्भसूची-ग्रंथ-

1. Abbi , Anvi ta (1994) : *Semantic Uni versal in Indian Languages*, Indi an Insti tute of Advance Study, Shi mla, Indi a.
2. Bhatt acharya, Pushpak (2004) : *Hindi Word Sense Di sambi guati on*, Inter nati onal Symposi um on MT, NLP and TSS, Del hi , Indi a.
3. Hockett Charles f. Indian Edition (1970) : *A Course In Modern Linguistic*, Oxford & IBH Publishing Comp.Pvt.Ltd.New Delhi , India.

4. Bharti, Akshar, Chaitanaya,Vineet & Sangal, Rajeev (2000) : *Natural Language Processing*, PHI, New Delhi, India.
5. Bharti, Akshar&Mannem Prashanth R.(2007) : *Introduction to the Shallow Parsing Contest For South Asian Languages*, Proceeding of the IJCAI-2007 Workshop on SPSAL-2007, LTRC, IIIT, Hyderabad, India.