AUCHITHYAM
(A Monthly Peer-Reviewed Online Telugu Research Journal)
Language & Literature | Linguistics | Arts & Humanities | Social Sciences
ISSN 2583-4797 (Online)
Publishing by: Dr. Rambhatla Parvatheeswara Sarma, Dr. RVRM Trust, 12-121/17, Flat No. 402, Satyanagar, Chinamusidiwada, Pendurthy, Visakhapatnam-531173. AP, India. +91 9247859580.

AUCHITHYAM | Volume-06 | Issue-08 | July 2025 | Peer-Reviewed | ISSN: 2583-4797

Article Link: https://www.auchithyam.com/2025/jul_2025/full_papers/july25_07.php

DOI: https://doi.org/10.5281/zenodo.16419374

7. భారతీయ భాషలకు ఎల్డీసిఐఎల్ టూల్స్ & అప్లికేషన్స్

డా. మోదుగు కాశింబాబు
రీసోర్స్ పర్సన్, ఎల్డీసిఐఎల్,
భారతీయ భాషాసంస్థ,
మైసూర్, కర్నాటక.
సెల్: +91 9908683093. Email: kasimtelugu@gmail.com

సమర్పణ (D.O.S): 13.05.2025 ఎంపిక (D.O.A): 28.06.2025 ప్రచురణ (D.O.P): 01.07.2025

వ్యాససంగ్రహం

భారతదేశం భాషా వైవిధ్యానికి నిలయం. వందలాది భాషలు, వేలాది మాండలికాలు భారతీయుల జీవనశైలిలో అంతర్భాగం. ఈ భాషలను సాంకేతికరంగంలో స్థిరపరచడం, భవిష్యత్తు తరాలకు అందించడం అత్యవసరం. ఈ సందర్భంలోనే ప్రపంచవ్యాప్తంగా ఆయా దేశభాషలకు సంబంధించి భాషాసాంకేతికత అందుబాటులోకి వస్తున్నది. మనదేశంలో కూడా భారతీయ భాషలకు ముఖ్య కేంద్రమైన భారతీభాషాసంస్థ, మైసూరు ఆధ్వర్యంలో రూపొందిన భారతీయ భాషానిధి సమాఖ్య (ఎల్డీసిఐఎల్) భాషాసాంకేతికాభివృద్ధి దృష్ట్యా అనేక సాధనాలు (టూల్స్), అనువర్తనాల (అప్లికేషన్స్) ను రూపొందించింది. అయితే ఇక్కడ ప్రపంచీకరణ నేపథ్యంలో భాషాసాంకేతికత రంగంలో తెలుగు, ఇతర భారతీయ భాషల స్థానాన్ని వివరిస్తూ, భాషిణి, ఎఐ4భారత్, ట్రిపుల్ ఐటి హైదరాబాద్, ఐఐటి మద్రాస్ వంటి భారతీయ సంస్థలకు, మైక్రోసాఫ్ట్ బింగ్, గూగుల్, కోపైలెట్, ఛాట్ జిపిటి వంటి విదేశీ సంస్థలకు దీటుగానూ, కొంత మెరుగ్గానూ రూపొందిన ఈ టూల్స్, అప్లికేషన్స్ పనితీరును, ఆధునిక అవసరాల దృష్ట్యా వాటి ఉపయోగాలను, భవిష్యత్ ప్రయోజనాలను వివరించే ప్రయత్నం చేశాను.

Keywords: భారతీయ భాషలు, భాషాభివృద్ధి సంస్థలు, భాషాసాంకేతికత, విదేశీ భాషాసాధనాలు, అనువర్తనాలు, భాషానిధి, కార్పస్, అనువాదం, కీబోర్డ్, ఇమేజ్, ఒసియార్, టెక్స్ట్, స్పీచ్

1. ప్రవేశిక

ప్రపంచభాషల్లో భారతీయభాషలది ఒక ప్రత్యేకస్థానం. ఇండో-ఆర్యన్, ద్రావిడ, ఆస్ట్రో-ఆసియాటిక్, టిబిటో-బర్మన్ వంటి నాలుగు రకాల భాషాకుటుంబాలకు చెందిన భాషలు మనదేశంలో ఉన్నాయి. ఇవి కేవలం సంభాషణలకు మాత్రమే పరిమితంకాక వేలసంవత్సరాలుగా గొప్ప సాహిత్యసంపదను, కళలు, సంస్కృతి, సంప్రదాయాలు, వివిధ శాస్త్రసంబంధ విషయాలను భద్రపరచి తరతరాలకు వారసత్వంగా అందిస్తున్న ఎంతో అభివృద్ధిచెందిన ప్రాచీన భాషలు. ప్రస్తుతం భాషాసామాజికరంగంలో అనేక విప్లవాత్మకమైన మార్పులు చోటుచేసుకుంటున్న సందర్భంలో కేవలం భాషల పరిశోధన, పరిరక్షణలతోపాటు సాంకేతికత ప్రధాన అవసరంగా పరిణమించింది. ఈ క్రమంలో భాషల డిజిటలీకరణ అత్యంత కీలకాంశమైంది. ఈ నేపథ్యంలో, భారతీయభాషలను కంప్యూటింగ్, ఇంటర్నెట్, మొబైల్, ఆర్టిఫిషియల్ ఇంటెలిజెన్స్ వంటి రంగాలలో సమర్థవంతంగా ఉపయోగించేందుకు అనేక ప్రయత్నాలు జరుగుతున్నాయి.

సహజ భాషా ప్రక్రియ (NLP), భాషావిశ్లేషణ, యాంత్రిక అనువాదం, స్వయంచాలక ప్రసంగ గుర్తింపు (ASR), పాఠ్యం నుంచి ప్రసంగంలోకి (TTS) వంటి సాంకేతికాభివృద్ధి వ్యవస్థల్లో భారతీయభాషలు తమదైన పాత్రను పోషించేందుకు సిద్ధమవుతున్నాయి. భారతప్రభుత్వం, కేంద్ర భాషాసంస్థలు, విద్యాసంస్థలు, ప్రైవేటురంగాలు భారతీయ భాషల సాంకేతికత అభివృద్ధికి కృషిచేస్తున్నాయి. ప్రధానంగా భాషిణి, ఎఐ4భారత్, సిఐఐఎల్, ఎల్డీసిఐయల్, ట్రిపుల్ ఐటి, టిడిఐఎల్ వంటి సంస్థలు భారతీయ భాషల్లో నూతన ప్రయోగాలు చేస్తున్నాయి. వీటి ఆధారంగా భవిష్యత్తులో భారతీయ భాషల్లో ఆధునిక సాంకేతిక అనువర్తనాలు మరింత విస్తృతమయ్యే అవకాశం ఉంది. ఈ సందర్భంలో భాషా సాంకేతికరంగంలో ప్రధానకేంద్రమైన భారతీయ భాషానిధి సమాఖ్య (లింగ్విస్టిక్ డేటా కన్సోర్టియమ్ ఫర్ ఇండియన్ లాంగ్వెజెస్-ఎల్డీసిఐయల్) ను గురించి, భాషాసాంకేతికతకు సంబంధించి ఇది సంకల్పించిన అనేక లక్ష్యాలను, అభివృద్ధిపరచిన సాధనాలను (టూల్స్), అనువర్తనాలను (అప్లికేషన్స్) గురించి తెలియజేయటం ఇక్కడి ప్రధానాంశం.

ఎల్డీసిఐయల్ మానవవనరుల అభివృద్ధి మంత్రిత్వశాఖ, ఉన్నత విద్యాశాఖ, భారతప్రభుత్వం ద్వారా 2007లో ఏర్పాటుచేయబడిన పథకం (స్కీమ్). ఇది భారతీయ భాషల సమగ్రాభివృద్ధికి భారతప్రభుత్వం ఏర్పాటుచేసిన 'భారతీయ భాషాసంస్థ'లో (సిఐఐయల్) భాగంగా ఉంది. ఎల్డీసిఐయల్ 2019 ఏప్రిల్ 4వ తేదీన భారత ఉపరాష్ట్రపతి శ్రీ వెంకయ్యనాయుడుచేత ప్రారంభించబడిన "ఎల్డీసిఐయల్ దత్తాంశ పంపిణీ వేదిక" (డేటా డిస్ట్రిబ్యూషన్ పోర్టల్) ద్వారా కృత్రిమమేధ (Artificial Intelligence-AI), సహజ భాషా ప్రక్రియ (Natural Language Processing) కోసం భాషా వనరులను అందించడం ప్రారంభించింది.

ఎల్డీసిఐయల్ లో భారతీయ భాషలైన సంస్కృతం, హిందీ, తెలుగు, బెంగాలీ, మరాఠీ, తమిళం, కన్నడం, సింధీ, బోడో, ఉర్దూ, గుజరాతీ, మైథిలీ, అస్సామీ, ఒడియా, మలయాళం, కొంకణీ, కశ్మీరీ, డోగ్రీ, పంజాబీ, మణిపురీ, నేపాలీ, సింధీ, సంతాలీ వంటి భాషలపై అనేక రకాలైన అధ్యయనాలు జరుగుతున్నాయి. ఈ భాషలకు సంబంధించి ఆయా రాష్ట్రాలలోని అర్హులైన భాషానిపుణులు ఇక్కడ పనిచేస్తున్నారు. భాషాసాంకేతిక రంగంలో నిపుణులైన డా. నారాయణ్ కుమార్ చౌదరి ప్రస్తుతం దీనికి అధిపతిగా వ్యవహరిస్తున్నారు. భారతీయ

భాషలలో అందుబాటులో ఉన్న పాఠ్యపుస్తకాలలో సాహిత్యం, సమాజశాస్త్రం, జంతు, వృక్షశాస్త్రాలు, గణితం, సాంఖ్యశాస్త్రం, వైద్యం, పర్యాటకం మొదలైన శాస్త్రరంగాలనుంచి నాణ్యమైన పాఠ్య వాగ్భాషానిధుల్ని తయారుచేసి, అభివృద్ధి చేయటం దీని ప్రధాన పని. భాషావిశ్లేషణ, ప్రక్రియల కోసం ఉపయోగించగల కొన్ని సాఫ్టువేర్ సంబంధిత పరిష్కారాలను ఇది అభివృద్ధి చేసింది. భాషాసాంకేతికతకు అవసరమైన సామగ్రిని అందించటంతోపాటు, వివిధ సాంకేతిక సాధనాలను (టూల్స్), అనువర్తనాలను (అప్లికేషన్స్) తయారుచేసి పరిశోధకులకు, పండితులకు అందుబాటులో ఉంచుతుంది. కొన్ని అప్లికేషన్‌లు అనువాద వ్యవస్థ, లిప్యంతరీకరణ, గ్రాఫీమ్ టు ఫోనీమ్ కన్వర్టర్లు, ట్రాన్స్క్రిప్షన్ వ్యవస్థ, కార్పస్ శోధన వ్యవస్థ, టెక్స్ట్ రీడర్లు, కీబోర్డ్, ఎడిట్ డిస్టెన్స్ కాలిక్యులేటర్, వర్డ్ ఫ్రీక్వెన్సీ కౌంటర్ వంటి వివిధ రకాల భాషానిధుల (Corpus) విశ్లేషణ, నిర్వహణ సాధనాలను అభివృద్ధి చేసింది. భాషాసాంకేతికతకు సంబంధించిన పలు వర్కుషాప్‌లు, శిక్షణ కార్యక్రమాలను, ఓరియంటేషన్ ప్రోగ్రామ్‌లను నిర్వహిస్తుంది. ఇక్కడ తయారైన నాణ్యమైన డేటా (దత్తాంశం)ను భాషాసాంకేతికరంగంలో కృషిచేస్తున్న ఆసక్తిగల డెవలపర్లకు, పరిశోధకులకు, సంస్థలకు పంపిణీ చేయాలని భావిస్తోంది. ప్రస్తుతం భారత ప్రభుత్వంచేత నిధులను పొందుతున్న ఈ సమాఖ్య తన సొంత నిధులను సమకూర్చుకుని, స్వయం సమృద్ధి కలిగిన సంస్థగా అవతరించాలని ప్రయత్నిస్తుంది.

ఎల్డీసిఐయల్ మేధా భాషిక అనే పేరుతో భారతీయ భాషలకు కృత్రిమ మేధ వేదికను రూపొందించింది. భారతీయ భాషలపై దృష్టి సారించే సహజ భాషా ప్రక్రియ, కృత్రిమ మేధస్సు, యంత్రప్రజ్ఞ (Machine Learning), పెద్ద భాషా నమూనాల (Large Language Models) వంటి రంగాలలో పనిచేయటానికి కావలసిన అనువర్తనాలను ఈ వేదిక రూపొందిస్తుంది. ఈ కృత్రిమమేధ సాధనాలను, అనువర్తనాలను 20-21 మార్చి, 2025 న 'ఏఐ బెంచ్ మార్క్' అనే అంశంపై నిర్వహించిన రెండు రోజుల జాతీయ సదస్సు వేదికగా ఆవిష్కరించింది.

2. కృత్రిమమేధ సాధనాలు

అనువాదిక: అనువాదిక భారతీయభాషాసంస్థకు చెందిన యంత్రానువాద సాధనం. ఇది అనువాద నమూనాలను, లిప్యంతరీకరణ ఇంజిన్లు, ASR ఇంజిన్లు మరియు TTS ఇంజిన్లను ఒకే ప్లాట్‌ఫామ్‌లో అందిస్తుంది. అంటే మూలపాఠాన్ని టైపింగ్ ద్వారా కానీ, రికార్డింగ్ ద్వారా కానీ, డాక్యుమెంట్ ద్వారా కానీ, వెబ్సైట్ లో ఉన్న పాఠాన్ని నేరుగా కానీ అందించవచ్చు. ఇది తాజాగా భాషిణి మోడళ్ల పై నిర్మించబడింది. వివిధ సమాంతర భాషానిధుల ఆధారంగా, ప్రధానంగా ఎల్డీసిఐఎల్ లో అందుబాటులో ఉన్న నాణ్యమైన పాఠ్యభాషానిధుల ఆధారంగా రూపొందించబడింది. ప్రత్యేకంగా ఈ వేదిక TDIL, NLTM, LDC-IL ఆధారంగా IIT మద్రాస్ లోని పరిశోధనా ప్రయోగశాల అయిన AI4Bharat నుండి IndicTrans2 (Gala et.al, 2023) ను ఉపయోగిస్తుంది. భాషిణి, ఏఐ4భారత్, గూగుల్, మైక్రోసాఫ్ట్ బింగ్ వంటి అనువాద వేదికే అయినా నిర్దుష్టమైన, నాణ్యమైన ప్రమాణాలు కలిగిన దత్తాంశసమితి ఆధారంగా రూపొందించబడింది. మీరు ఏదైనా భారతీయ భాషలోని వచనాన్ని (ఇంగ్లీషుతో సహా) ఇచ్చి, ఏ భారతీయ భాషలోనైనా (ఇంగ్లీషుతో సహా)

అనువాదాన్ని అడిగి పొందవచ్చు. ఇందులో ఆంగ్లం, హిందీ, తెలుగు, తమిళం, కన్నడ, మలయాళం, సంస్కృతం, బెంగాలీ, అస్సామీ, అవధీ, భోజ్ పురి, గుజరాతీ, ఛత్తీస్ గరీ, కాశ్మీరీ, మైథిలీ, మరాఠీ, మీతై, నేపాలీ, ఒడియా, పంజాబీ, సంతాలీ, సింధీ, ఉర్దూ వంటి మొత్తం 23 భాషలకు అనువాదం అందించే సామర్థ్యం ఉంది. అయితే, వినియోగదారులు ఇప్పటికీ వారు అనువదించాలనుకుంటున్న భాషను ఎంచుకోవచ్చు. మూలపాఠం, లక్ష్యపాఠం రెండింటి లిప్యంతరీకరణను కూడా ఇక్కడ పొందవచ్చు. ప్రాప్యత (access) సౌలభ్యం కోసం, మూలపాఠంలోని భాషను స్వతస్సిద్ధంగా గుర్తించడానికి లాంగ్వేజ్ డిటెక్షన్ మాడ్యూల్స్ కూడా అందుబాటులో ఉన్నాయి. ఈ సాధనానికి గూగుల్ ట్రాన్స్లేషన్ అంతటి ప్రాచుర్యం లేదు, కానీ నాణ్యమైన అనువాదంకోసం దీనిని ఎంచుకోవటం మంచిది. https://anuvadika.ciil.org/ అనే జాలవేదిక ద్వారా ఉపయోగించవచ్చు.

లిప్యంతర: లిప్యంతర అనేది పాఠాన్ని ఒక లిపి నుండి మరొక లిపికి మార్చే లిప్యంతరీకరణ అనువర్తనం. ఆంగ్లం, తెలుగు, హిందీ, తమిళం, కన్నడ, మలయాళం, బెంగాలీ, ఒడియా, పంజాబీ, గుజరాతీ, ఉర్దూ వంటి పది లిపులు ఇక్కడ అందుబాటులో ఉన్నాయి. దీనికి సంబంధించి ఇందులో యూనికోడ్ క్యారెక్టర్ మ్యాపింగ్ యంత్రాంగం ఉపయోగించబడింది. ఇది బ్రాహ్మీ ఆధారిత లిపులను ఉపయోగించే అన్ని భారతీయ భాషలకు బాగా పనిచేస్తుంది (అరబిక్ లిపిలో ఉన్న ఉర్దూ, కాశ్మీరీ భాషలు మినహా అన్ని భారతీయ భాషలకు ఉపకరిస్తుంది). రోమన్ లిపిలోకి లిప్యంతరీకరణ కోసం, అరబిక్ లిపిలో ఉన్న ఉర్దూ, కాశ్మీరీలకు ఉపకరించటానికి ఇండిక్ ట్రాన్సుమోడల్‌ను ఉపయోగిస్తున్నారు. (Bhat et. al, 2015). ప్రస్తుతం అరబిక్ లిపి నుండి (ఉర్దూ/కాశ్మీరీ) లిప్యంతరీకరణ అందుబాటులో లేదు. దీనిని https://lipyantara.ldcil.org/ మీద నొక్కడం ద్వారా చేరుకోవచ్చు.

లిపిద: లిపిద అనేది ఎల్డీసిఐఎల్ అభివృద్ధి చేసిన వెబ్ ఆధారిత ఆప్టికల్ క్యారెక్టర్ రికగ్నిషన్ (ఒసిఆర్) సాధనం. పరిశోధనలో భాగంగా వివిధ సంస్థలు, వ్యక్తులు అభివృద్ధి చేసిన వైవిధ్యమైన ఒసిఆర్ నమూనాలను అనుసంధానిస్తుంది. ఈ సాధనంలో అందుబాటులో ఉన్న ఒకే వేదికపై పనిచేయడానికి అవన్నీ కలిసి తీసుకోబడ్డాయి. బేస్ మీద, మేము టెసెరాక్ట్, బేర్-బోన్స్ ఒసిఆర్ ఇంజిన్ను ఉపయోగిస్తాము, ఇది ఓపెన్-సోర్స్ ఒసిఆర్ ఇంజిన్. మేము దానిని వెబ్ అప్లికేషన్ గా మార్చడానికి దాని పైన ఒక చుట్టను నిర్మిస్తాము. ఏదైనా వెబ్ అప్లికేషన్లో ఉపయోగించగల మరియు ఇన్పుట్ను. పిడిఎఫ్ లేదా. జెపిఇజి చిత్రంగా ఇన్పుట్గా తీసుకునే ఎపిఐని కూడా నిర్మిస్తాము. వినియోగదారులు చిత్రం ఉన్న భాష/లిపిని కూడా ఎంచుకోవాలి. చిత్రంలోని వచనం బహుళ భాషలలో ఉంటే ఇది బహుళ భాషలను కూడా గుర్తించగలదు. ఇది https://lipidha.ldcil.org/ ద్వారా అందుబాటులో ఉంది.

అనులేఖిక: అనులేఖిక అనేది అన్ని భారతీయ భాషలకు పనిచేసే వెబ్ ఆధారిత ఆటోమేటిక్ స్పీచ్ రికగ్నిషన్ వ్యవస్థ ఆధారంగా రూపొందించిన ఒక ట్రాన్స్క్రిప్షన్ సాధనం. ఈ వెబ్ అనువర్తనాన్ని ఎల్డీసిఐఎల్ టీమ్ అభివృద్ధి చేసింది. ఈ టూల్ వినియోగదారులను నేరుగా మాట్లాడటానికి, ముందుగానే రికార్డు చేసిన ఆడియో

ఫైల్‌ను అప్‌లోడ్ చేయడానికి అనుమతిస్తుంది. అందించిన సంబంధిత పాఠాన్ని కావలసిన భారతీయ భాషలోకి లిప్యంతరీకరించి అందిస్తుంది. ప్రస్తుత ASR మోడల్‌ను ఫేస్ బుక్/మేట (Pratap et.al) నుండి మాసివ్లీ మల్టీలింగ్యువల్ స్పీచ్ (MMS) ప్రాజెక్ట్ నుండి సేకరించారు. ఈ మోడల్ ఆర్కిటెక్చర్ ASR మోడల్ యొక్క ప్రధాన ఆర్కిటెక్చర్‌కు క్లిప్ చేసే మాడ్యులైజ్డ్ అడాప్టర్ మాడ్యూల్‌లను ఉపయోగిస్తుంది. ఇది మరింత బలమైన, ఖచ్చితమైన బహుభాషా పనితీరును అందిస్తుంది. ఇది ASR మోడల్ కింద హిందీ, తెలుగు, తమిళం, కన్నడ, మలయాళం, అవధి, మరాఠీ, బోడో పర్గా, ఇంగ్లీష్, అస్సామీ, బెంగాలీ, గుజరాతీ, హర్యాన్వి, మణిపురి, మార్వారీ, సింధీ వంటి 16 భారతీయ భాషలలో ఈ సౌకర్యాన్ని అందిస్తుంది. ఇవికాక మైథిలి ASR వ్యవస్థ ఇక్కడ అందుబాటులో ఉన్న ప్రత్యేక ఇంటర్‌ఫేస్.

ఈ టూల్ పరిమిత వనరులతో కూడిన వ్యవస్థలో హోస్ట్ చేయబడింది. అందువల్ల, ఇది 50MB వరకు మాత్రమే ఆడియో ఇన్‌పుట్ ఫైల్‌ను తీసుకుంటోంది. అందువల్ల ఒకేసారి ఒక నిమిషం కంటే ఎక్కువ ఆడియోను రికార్డ్ చేయకూడదు, ఎందుకంటే పరిమాణం 50MB కంటే ఎక్కువగా ఉండవచ్చు లేదా ప్రాసెసింగ్, అప్‌లోడ్ చేయడంలో సమస్యలు ఎదురయే అవకాశం ఉంది. ఈ టూల్ విరామ చిహ్నాలను గుర్తించలేదు. దీనిని https://anulekhika.ldcil.org/ పై నొక్కడం ద్వారా చేరుకోవచ్చు.

అనువాచిక: అనువాచిక అనేది వచనాన్ని ప్రసంగం (Text to Speech)లోకి మర్చగలిగే సాధనం. ఇది ఇప్పటికే అందుబాటులో ఉన్న నమూనాల ఆధారంగా రూపొందించబడింది. ఈ సాధనం దానికి ఇచ్చిన ఏ వచనాన్ని అయినా మాట్లాడగలదు. ఇది ప్రస్తుతం హిందీ, తెలుగు, కన్నడం, తమిళం, మలయాళం, బెంగాలీ, గుజరాతీ, ఒడియా, అస్సామీ, రాజస్థానీ, ఇంగ్లీష్ భాషలకు మాత్రమే అందుబాటులో ఉంది. ఇతర భాషలకు సంబంధించిన అనేక వాగ్దత్తాంశాలను (Speech Data) రూపొందిస్తున్నది. తద్వారా ఇంకా ఇతర భారతీయ భాషలను కూడా జోడించి మరింత మెరుగ్గా నవీకరించే దిశగా కృషి చేస్తున్నారు. https://anuvachika.ldcil.org/ ఈ వెబ్ సైట్ ద్వారా దీనిని పరిశీలించవచ్చు.

3. ఇతర సాధనాలు (టూల్స్)

శబ్ద సంధాన్: శబ్ద్ సంధాన్ అనేది ఎల్డీసిఐఎల్ లో అందుబాటులో ఉన్న పాఠ్యదత్తాంశం (Text Corpus), వాగ్దత్తాంశాలను (Speech Corpus) పరిశీలించే దత్తాంశశోధన సాధనం. ఇది ప్రత్యేకంగా ఒకటి లేదా అంతకంటే ఎక్కువ భాషలపై పరిశోధన చేయాలనుకునే భాషావేత్తలకు, పరిశోధకులకు అనుకూలిస్తుంది. ఇది బహుభాషా దత్తాంశ శోధన. ఇక్కడ కావలసిన ఒక పదం కోసం శోధించినప్పుడు ఆ పదంయొక్క ఫ్రీక్వెన్సీని చూడవచ్చు, ఆ పదాలను కలిగి ఉన్న వాక్యాలను కూడా కనుగొనవచ్చు. ఇది ఒక రకంగా మన పదప్రయోగకోశాలను గుర్తుచేస్తుంది. ఇంకా ఈ సందర్భంలో భారతీయ భాషలన్నిటిలో అందుబాటులో ఉన్న పదాల ఉచ్చారణలు కూడా వినవచ్చు. దీని వెబ్సైట్ ని ఇక్కడ గమనించవచ్చు.
(https://shabd.ldcil.org)

ధ్వని పరివర్తక్: ధ్వని పరివర్తక్ (Media Converter) అనేది ఆడియో ఫైలుని మార్చే సాధనం. మనం అందించే ఆడియోను ఒక రూపం (Format) నుండి మరొక రూపానికి మారుస్తుంది. వీడియో, ఆడియో, ఇతర మల్టీమీడియా ఫైళ్లు, స్ట్రీమ్లను నిర్వహించడానికి లైబ్రరీలు, ప్రోగ్రామ్ ల సూట్ ను కలిగి ఉన్న ఉచిత మరియు ఓపెన్ సోర్స్ సాఫ్ట్వేర్ ప్రాజెక్ట్ అయిన ఎఫ్ఎఫ్ఎంపెగ్ పైన కన్వర్టర్ అభివృద్ధి చేయబడింది. వినియోగదారులు ఆడియో ఫైళ్ళను ఒక ఫార్మాట్ నుండి మరొక ఫార్మాట్ కు మార్చవచ్చు. ఇది. వావ్,. ఎం. పి. 3,. వెబ్మ్ మొదలైన అనేక ఆడియో ఫైల్ ఫార్మాట్లకు మద్దతు ఇస్తుంది. https://dhvani.ldcil.org/

అక్షరాంక: 'అక్షరాంక' అనేది పదంలోని వర్ణభాగాలను వేరుచేసి ఇచ్చేది. మనం సూచించే పదాలలోని అక్షరాలను (Character Split) విడదీస్తూ, ప్రతి అక్షరం యొక్క దశాంశ (Unicode codepoint value in Decimal), హెక్సా-దశాంశాలలో (Unicode codepoint value in Hexadecimal) యూనికోడ్ విలువలను అందిస్తుంది. మనకు తెలుసు కంప్యూటర్లు అక్షరాలతోకాక సంఖ్యలతోనే వ్యవహరిస్తాయి. కాబట్టి ప్రతి భాషలోని ఒక్కొక్క అక్షరానికి ఒక్కొక్క సంఖ్యను కేటాయిస్తుంది. దీనినే కంప్యూటర్ భాషలో యూనికోడ్ వాల్యూ అంటారు. కంప్యూటర్లు భాషకు ప్రోగ్రామింగ్ రాయాలంటే ఆ భాషలో ఉన్న అక్షరాలకు అవి ఇచ్చుకున్న యూనికోడ్ విలువే ఆధారం. పాఠ్యాన్ని క్రమబద్ధీకరించే సమయంలో సారూప్యంగా కనిపించి, కోడ్ పాయింట్లలో భిన్నంగా ఉండే అక్షరాలకు, చిహ్నాల మధ్య తేడాను గుర్తించి వాటి యూనికోడ్ విలువల ఆధారంగా ప్రోగ్రామ్ లను రూపొందించడానికి ప్రోగ్రామర్లకు ఇది సహాయపడుతుంది. దీనిని ఉపయోగించటానికి ఈ క్రింద ఇచ్చిన వెబ్ సైట్ మీద నొక్కాలి.

https://medha.ciil.org/en/WebApplication/aksharanka

నుడియలవి: 'నుడియలవి' (Audio Duration and Size Calculator) అనేది ఒక ఆడియో ఫైల్ కాలవ్యవధిని, పరిమాణాన్ని కొలిచే సాధనం. ఇది wav, mp3, aac, wma ఫార్మాట్లలో ఉన్న ఆడియో ఫైళ్ళవ్యవధి, పరిమాణాన్ని సంగ్రహించి పట్టిక చేస్తుంది. ఇది బ్యాచ్ ప్రాసెసింగ్ ద్వారా నిర్వహిస్తుంది. అంటే పెద్ద మొత్తంలో ఉన్న డేటాను వేగవంతంగా ప్రాసెస్ చేయటానికి డేటాను భాగాలుగా విభజించుకుంటుంది. ఒకే ఆపరేషన్లో అనేక ఫైళ్ళ విశ్లేషణకు వీలు కల్పిస్తుంది. ఈ టూల్ అన్ని ప్రాసెస్ చేయబడిన ఆడియో ఫైళ్ళ మొత్తం వ్యవధిని, మొత్తం పరిమాణాన్ని కూడా లెక్కించి ఇస్తుంది. నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్, స్పీచ్ టెక్నాలజీలపై పనిచేస్తున్న వారికి స్పీచ్ కార్పస్ పరిమాణం, వ్యవధిని తెలుసుకోవడంలో ఈ అప్లికేషన్ ప్రత్యేకంగా సహాయపడుతుంది. దీనిని ఈ వెబ్ సైట్ https://medha.ciil.org/en/WebApplication/nudiyalavi ద్వారా పొందవచ్చు.

పాఠాంతర: పాఠాంతర (Levenshtein Distance Calculator) ఎల్డీసిఐఎల్ వెబ్ అప్లికేషన్. లెవెన్స్టెయిన్ డిస్టెన్స్ అనేది రెండు వరుసల మధ్య వ్యత్యాసాన్ని కొలిచే ఒక మాత్రిక. ఇది ఒక పదాన్ని మరొక పదంగా మార్చడానికి అవసరమైన కనీస ఏకాక్షర సవరణలు అంటే ఒక అక్షరాన్ని చేర్చడం, తొలగించడం లేదా ప్రత్యామ్నాయాన్నివ్వడం వంటి వాటి సంఖ్యను లెక్కించే ఆల్గొరిథం. ఈ ఆల్గొరిథం సమాచార సిద్ధాంతం,

భాషాశాస్త్రం, కంప్యూటర్ సైన్స్ వంటి వివిధ రంగాలలో విస్తృతంగా ఉపయోగించబడుతుంది. కవులు, రచయితలు, సంపాదకులు ఒక పాఠానికి సంబంధించిన వివిధ పాఠాంతరాల మధ్య సారూప్యతను అంచనా వేయడానికి ఈ పాఠాంతర టూల్ ను ఉపయోగించవచ్చు. స్పెల్ చెకింగ్, ఒసిఆర్, ఎఎస్ఆర్ వ్యవస్థల నాణ్యత నియంత్రణ కోసం వాటి ఫలితాన్ని సరిదిద్దిన తరువాత డేటాతో పోల్చుకోవటానికి ఇది ఉపయోగపడుతుంది. ఇచ్చిన పాఠాంతరాల మధ్య సవరించిన దూరాన్ని లెక్కిస్తుంది. పాఠ సవరణ రేటును లెక్కించడానికి ఉపయోగపడుతుంది. పదస్థాయిలో సవరణ రేటును లెక్కించడానికి, సవరణలను టోకనైజ్ చేయటానికి ఉపయోగపడుతుంది. క్రింది వెబ్ సైట్ ద్వారా ఈ టూల్ ను ఉపయోగించవచ్చు.
(https://medha.ciil.org/en/WebApplication/paataantara)

కణజ: కణజ అనేది ఎల్డీసిఐఎల్ పాఠ్యదత్తాంశాన్ని నిక్షిప్తం చేయటానికి రూపొందించిన ప్రత్యేక సాధనం. ఇది కన్నడ పదం. దీనికి కన్నడంలో ధాన్యాన్ని సంగ్రహించే నేలమాలిగ అని అర్థం. అంటే ఇదొక గోడౌన్ లాంటిది. కర్నాటక ప్రభుత్వం అధికారికంగా కణజ అనే ఒక పోర్టల్ ద్వారా వివిధ సాహిత్య ప్రక్రియల్లో ఉన్న కన్నడ సాహిత్యాన్నంతా ఒకచోట చేర్చి, అంతర్జాలంలో అందుబాటులో ఉంచింది. ఈ సందర్భాన్ని దృష్టిలో ఉంచుకొని కర్నాటక మైసూర్లో ఉన్న భారతీయ భాషాసంస్థలో భాగమైన ఎల్డీసిఐఎల్ తన పాఠ్యదత్తాంశానికి కణజ అనే పేరుతో భారతీయ భాషల్లో డేటాను సేకరించి ఇక్కడ నిక్షిప్తం చేస్తుంది.

ఇది కేవలం ఎల్డీసిఐఎల్ కి సంబంధించింది కావటంవల్ల డేటా రక్షణ నిమిత్తం ఇక్కడ పనిచేసేవారికి మాత్రమే డేటా యాక్సెస్ ఇవ్వబడుతుంది. దాదాపు 28 భారతీయభాషల్లో పాఠ్యదత్తాంశం ఇక్కడ అందుబాటులో ఉంది. ఈ భారతీయ భాషల్లో సాహిత్యం, గణితం, సమాజశాస్త్రం, విద్య, వైద్యం, జీవశాస్త్రం, సాంఖ్యకశాస్త్రం వంటి భిన్నరంగాలలో అందువాటులో ఉన్న ప్రభుత్వ పాఠ్యపుస్తకాలు, వివిధ కవుల, రచయితల పుస్తకాలు, వార్తాపత్రికలు, వివిధ మాసపత్రికల్లో లభిస్తున్న పాఠ్యదత్తాంశాన్ని అనుమతి ద్వారా తీసుకుంటారు. ఆ పుస్తకాలను ఓసియార్ ద్వారా పిడియఫ్ నుంచి వర్డ్ ఫార్మెట్ కు మార్చి దానిని కణజలోకి ఎక్కిస్తారు. ఈ సందర్భంలో ఒక్కొక్క పుస్తకానికి ఒక్కొక్క ఫైల్ వరుససంఖ్య, పుస్తకంపేరు, పాఠ్యం విభాగం, ఉపవిభాగం, రచయిత, ముద్రణాలయం, ముద్రణ ఊరు, సంవత్సరం వంటి పలు గుర్తులను నమోదు చేస్తారు. దీనిని మెటాడేటా అంటారు. ఆ తరువాత ఇక్కడ రెండో విభాగం కంటెంట్. ఈ కంటెంట్ లో పుస్తకంలోని కొంతభాగాన్ని (సాంపుల్స్) తీసుకొని అక్కడ ఉంచుతారు. ఈ కణజలోనే నాణ్యత, ప్రమాణాల దృష్ట్యా విషయనిపుణుల చేత అక్షరదోషాలు, వ్యాకరణదోషాలు లేకుండా నాణ్యమైన దత్తాంశంగా తయారుచేస్తారు. దీనిని సేవ్ చేయటం ద్వారా ఇది కణజలో భద్రపరచబడుతుంది. ఈ పాఠ్యానికి సంబంధించిన మొత్తం పదాలసంఖ్య, ఒక పదం ఎన్నిసార్లు పునరావృతమైన విషయం, పాఠ్యం పునరావృతమైన (డూప్లికేషన్) విషయం వంటివి దత్తాంశం నాణ్యతను పెంచేందుకు రూపొందించినవి. ఇక్కడ నమోదు చేసిన వివరాలను వివిధ విషయాల ఆధారంగా ప్రత్యేకంగా చూడవచ్చు. ఈ టూల్ లో భద్రపరచిన భారతీయ భాషల పాఠ్యదత్తాంశాన్ని

ఎల్డీసిఐఎల్ పరిశోధకులకు నియమనిబంధనల ద్వారా ఉచితంగా అందుబాటులో ఉంచుతుంది. ప్రైవేట్ భాషా సంస్థలకు రుసుము ద్వారా అందిస్తుంది.

ట్రాన్సిట్ టూల్: అనువాదం ఆధారంగా భారతీయ భాషలన్నిటికి భాషాసాంకేతికతను అందించే లక్ష్యంతో రూపొందించిన ప్రతేకమైన సాధనం ట్రాన్సిట్. భారతదేశంలోని 2011 జనాభా లెక్కల ప్రకారం నిర్ణయించిన 270 మాతృభాషలకు భాషా వనరులను అభివృద్ధి చేయడానికి రూపొందించిన సమాంతర పాఠ్యదత్తాంశ యోజన (Parallel Corpora Project). ఈ ప్రాజెక్ట్ ద్వారా అయా భాషలకు సమాంతర కార్పస్ ను సృష్టించడం ప్రధాన లక్ష్యం. తద్వారా సాంకేతికాభివృద్ధి రంగంలో షెడ్యూల్డ్ భాషలతో పాటు మైనారిటీ భాషలు కూడా అభివృద్ధి చెందుతాయి. ప్రధానంగా సాంకేతికత రంగంలో అడుగుపెడతాయి. తద్వారా ఈ భాషల ఉనికి కాపాడబడుతుంది. ఈ పని ద్వారా భాషావైవిధ్యాన్ని పరిరక్షించడం, అధ్యయనం చేయడం, ప్రోత్సహించడం వంటి ప్రధాన లక్ష్యాలు నెరవేరుతాయి. ఈ ప్రాజెక్ట్ భారతీయ భాషా వారసత్వ పరంపరను నిలుపుతుంది. ఈ భాషలకు సాంస్కృతిక సంరక్షణతోపాటు విద్యా వనరులకు అవకాశం ఏర్పడుతుంది. భారతప్రభుత్వం అధికారికంగా గుర్తించిన 22 భాషలతోపాటు ఆయా రాష్ట్రాల్లో వివిధ జనసముదాయాల అసిత్వంలో ఉన్న భాషలు ఇందులో ఉన్నాయి. ఉదాహరణకు తెలుగురాష్ట్రంలో మాట్లాడే గోండి, కొలామీ, లంబాడి, సుగాలి, కొండ, కోయ, గదబ, కోదు, ఎరుకల వంటి భాషలు. ఇలా తమిళం, కన్నడం వంటి రాష్ట్రాల్లో ఉన్న భాషలు, ఉత్తరాదిలో ఉన్న భాషలు, ఈశాన్యరాష్ట్రాల్లో ఉన్న అనేక గిరిజన భాషలు కలిపి మొత్తం 270 భాషలలో ఈ పని జరుగుతుంది. అయితే ఈ భాషలకు ఆయా రాష్ట్రాల్లో ఉన్న ప్రధానభాషలు మూలభాషగా ఉంటాయి. ఉదాహరణకు తెలుగు రాష్ట్రాల్లో ఉన్న భాషలకు తెలుగును మూలభాషగా తీసుకోవచ్చు లేదా ఆంగ్లం, హిందీ భాషలను మూలభాషగా తీసుకోవచ్చు. ఈ మూలభాషనుంచి లక్ష్యభాషలకు అనువాదాలు ఈ టూల్ ద్వారా నిర్వహించబడతాయి.

ఈ టూల్ లో పనిచేయటానికి పైన తెలిపిన 270 భాషల్లో ఏదో ఒక భాషకోసం పనిచేయాలి. అందుకోసం మన వ్యక్తిగత వివరాలను నమోదుచేసుకోవాలి. వారు పంపిన లింక్ ద్వారానే మనకు ప్రవేశం దొరుకుతుంది. ఆ తరువాత మూలభాషకు సమాంతరమైన లక్ష్యభాషను ఆ టూల్ లోనే టైప్ చేయవలసి ఉంటుంది. ఈ అనువాదం ఇతరులచేత సమీక్షించబడుతుంది. ఈ టూల్ ఉపయోగానికి సంబంధించిన సూచనలన్నీ అందులో ఇవ్వబడతాయి. ఈ టూల్ కేవలం అనువాదానికి మాత్రమే పరిమితమైంది కాదు. పిడియఫ్ నుంచి వర్డ్ కన్వర్టర్, ఒసియార్, పాఠ్యదత్తాంశ ప్రూఫ్ రీడింగ్, డీజిటైజేషన్ వంటి అనేక పనులకు ఉపయోగించవచ్చు.

4. డెస్క్టాప్ అప్లికేషన్లు

'చాను': చాను (Meetei Mayek InScript) మణిపురి భాషకు సంబంధించిన ఇన్స్క్రిప్ట్ కీబోర్డ్. ఇది మీటీ మాయెక్ స్క్రిప్ట్ను అందిస్తుంది. ఇన్స్టాలేషన్ తర్వాత కీబోర్డ్ యుఎస్-ఇంగ్లీష్ కింద టాస్క్ బార్ యొక్క

లాంగ్వేజ్ మెనూలో అందుబాటులో ఉంటుంది. ఇది ఎం. ఎస్. కె. ఎల్. సి. అనువర్తనాన్ని ఉపయోగించి అభివృద్ధి చేయబడింది.

దర్పణ: దర్పణ తెలుగు, కన్నడ, మలయాళం లిపులను ఇంటర్నేషనల్ ఫోనెటిక్ ఆల్ఫాబెట్ (IPA) గా మార్చే ఎల్డీసిఐఎల్ టూల్. ఇది సంబంధిత లిపులలోని యూనికోడ్ 15.1 లో సంకేతనిర్మితమైన (Encode) అన్ని అక్షరాలను కవర్ చేస్తుంది. ఇది ఈ భాషల కోడ్ బ్లాక్ ల సంఖ్యలను ఇండో-అరబిక్ సంఖ్యలుగా కూడా మారుస్తుంది. అలాగే ఈ లిపి బ్లాక్ లకు చెందిన సాంస్కృతిక చిహ్నాలను మాత్రం అలాగే ఉంచుతుంది. భాషా సంప్రదాయాలను అనుసరించి ఇతర అక్షరాలనన్నీ ధ్వన్యాత్మక వర్ణమాలగా మార్చడానికి ప్రాసెస్ చేయబడతాయి. దీనిని లిపి నుంచి వర్ణం (Grapheme to Phoneme System) వ్యవస్థలు, స్పీచ్ రికగ్నిషన్, టెక్స్ట్-టు-స్పీచ్ (టిటిఎస్) వ్యవస్థలు, భాషా అభ్యాస అనువర్తనాలు, భాషావిశ్లేషణ సాంకేతికతలలో ఉపయోగించవచ్చు. ఇది ఈ వెబ్ సైట్ ద్వారా అందుబాటులో ఉంటుంది.
(https://medha.ciil.org/en/DesktopApplication/darpana)

లీమరెన్: లీమరెన్ అనేది ఎల్డీసిఐఎల్ అభివృద్ధి చేసిన ఒక ఫొనెటిక్ కీబోర్డ్. ఇది మీటీ మాయెక్ (మణిపురి) లిపికోసం తయారుచేయబడింది. ఇన్‌స్టాలేషన్ తర్వాత కీబోర్డ్ టాస్కుబార్‌లోని భాషా మెనూలో UK-ఇంగ్లీష్ కింద అందుబాటులో ఉంటుంది. ఇది MSKLC అప్లికేషన్ ఉపయోగించి అభివృద్ధి చేయబడింది. ఈ క్రింది లింకు ద్వారా దీనిని ఉపయోగించవచ్చు.
(https://medha.ciil.org/en/DesktopApplication/leimaren)

మిలి: మిలి ఎల్డీసిఐఎల్ రూపొందించిన ట్రాన్సిలిటరేటర్ అప్లికేషన్. ఇది వివిధ భారతీయ, రోమన్ లిపుల మధ్య మార్పిడిని సులభతరం చేయడానికి వద్ద అభివృద్ధి చేయబడిన లిప్యంతరీకరణ అనువర్తనం.

(అ) భారతీయ లిపి - రోమన్ లిపి మార్పిడి: మిలి పూర్వనిర్ధారిత స్కీమాను ఉపయోగించి భారతీయ లిపులను రోమన్ లిపికి, లిప్యంతరీకరణ చేస్తుంది. స్కీమా అనేది కంప్యూటర్ ప్రోగ్రామింగ్‌లో డేటాబేస్ నిర్మాణం.

(ఆ) భారతీయ లిపి - భారతీయ లిపి మార్పిడి: యూనికోడ్ ప్రమాణం ఉన్న భారతీయ లిపుల (తెలుగు, తమిళం, కన్నడ, మలయాళం, ఒడియా, గుజరాతీ, గురుముఖి, బెంగాలీ, దేవనాగరి) అక్షరాలను 128 కోడ్ పాయింట్ల గుణకాలతో సమానమైన అక్షరాల అంతరంతో నిర్వహిస్తుంది. ఈ 9 లిపులలో ప్రతి ఒక్కటి U + 0900 నుండి U + 0D7F పరిధిలో 128 కోడ్ పాయింట్ల బ్లాక్ ను కలిగివుంటుంది. మిలి ఈ లిపుల మధ్య లిప్యంతరీకరణ కోసం ఈ ప్రత్యేకమైన ఏర్పాటును ఉపయోగించుకుంటుంది.

ఉదాహరణలు: మలయాళ అక్షరం 'അ' (/a/) ను కన్నడకు లిప్యంతరీకరణ చేయడానికి 'హ' యొక్క యూనికోడ్ విలువ U + 0D05 = 3333 (దశాంశం); 'ಅ' యొక్క యూనికోడ్ విలువ U + 0C85 = 3205 (దశాంశం); కన్నడ, మలయాళ లిపి బ్లాక్ 1 బ్లాక్ వేరుగా ఉన్నందున, కన్నడ బ్లాక్ మలయాళ బ్లాక్ కు ముందు ఉన్నందున దానిని ఇలా లెక్కిస్తారు. టార్గెట్ అక్షరం = 3333 + (-1 * 128) = 3205, ఫలితంగా కన్నడ 'ಅ' అవుతుంది.

బ్రాహ్మి, చక్మా, కైతి, ఖరోష్టి, లెప్చా, లింబు, సౌరాష్ట్ర, శారదా, సిలోటి నాగరి, తాక్రి, తిర్హుటా వంటి లిపులకు కూడా అనుకూలిస్తుంది. ఈ లిపుల అక్షరాలకు లిప్యంతరీకరణ సూత్రాన్ని వర్తింపజేసే ముందు మధ్యవర్తిగా కన్నడకు మ్యాప్ చేయబడతాయి. అలాగే మీతై మయేక్ (మణిపురి) కోసం బంగ్లా లిపిని మధ్యవర్తిగా ఉపయోగిస్తారు. భాషాశాస్త్రవేత్తలు, పరిశోధకులు, బహుళ భారతీయ లిపులతో పనిచేసే వారికి ఈ 'మిలి' ప్రధానమైన సాధనం. దీనిని ఈ లింక్ ద్వారా చూడవచ్చు.
(https://medha.ciil.org/en/DesktopApplication/mili)

సీమ: సీమ (Iterative type-token analyser) పాఠ్య దత్తాంశానికి సంబంధించిన టైప్-టోకెన్ ను పునరావృతాన్ని విశ్లేషించే సాధనం. ఇందులో పాఠ్యదత్తాంశం XML లేదా TXT ఫార్మాట్లలో ఉండాలి. ప్రతి పునరావృతానికి పెరుగుదల రేటును లెక్కిస్తుంది. సీమ ప్రత్యేకంగా వినియోగదారు అందించిన లిపిలోని టోకెన్లను పరిగణిస్తుంది. సంఖ్యా టోకెన్లను మినహాయిస్తుంది, విరామ చిహ్నాలను వదిలేస్తుంది.

టైప్-టోకెన్ విశ్లేషణలో వివిధ పాఠాలు, రచనాశైలులు లేదా కాలవ్యవధులలో, భాషా వినియోగంలో తేడాలను తెలియజేస్తుంది. వివిధ దత్తాంశాల టైప్-టోకెన్ నిష్పత్తులను పోల్చడం ద్వారా పరిశోధకులు భాషాధోరణులను, శైలీకృత తేడాలను లేదా కాలక్రమేణ భాషలో జరుగుతున్న మార్పులను గుర్తించగలరు. కార్పస్‌లోని లెక్సికల్ అంశాలను కూడా ఈ విశ్లేషణ ద్వారా అంచనా వేయవచ్చు.

ఇది పరిశోధకులకు ప్రత్యేకమైన పదాలు (రకాలు), మొత్తం పదాలు (టోకెన్లు) రెండింటి లెక్కలను అందించడం ద్వారా కార్పస్‌లోని పదజాలం గొప్పతనంతోపాటు పదజాలంలోని వైవిధ్యాన్ని అర్థం చేసుకోవడానికి ఉపయోగపడుతుంది. పదజాల పరిమాణం, లెక్సికల్ వైవిధ్యం, పదవినియోగ నమూనాలను అంచనా వేయడానికి, పరిమాణాత్మక భాషా విశ్లేషణకు, పాఠ్యదత్తాంశంలోని భాషావినియోగానికి సంబంధించిన సంక్లిష్టతలను అర్థం చేసుకోవడానికి ఇది సరిగా ఉపయోగపడుతుంది.
(https://medha.ciil.org/en/DesktopApplication/seema)

క్విక్ ఫైల్ రీనేమర్: క్విక్ ఫైల్ రీనేమర్ (Bulk file renaming application) పెద్ద ఫైళ్లను పెద్దమొత్తంలో పేరు మార్చే ప్రక్రియను క్రమబద్ధీకరించడానికి రూపొందించబడిన ఎల్డీసిఐఎల్ అనువర్తనం. ఇది యూజర్లు నిర్వచించిన ప్రమాణాల ఆధారంగా ఫైల్ పేర్లను సవరించడానికి, ఉపయోగించడానికి సులభంగా, విస్తృత శ్రేణి ఫైల్ నిర్వహణ పనుల కోసం సమగ్రమైన లక్షణాలను అందిస్తుంది. ఇందులో యూజర్ ఇంటర్ఫేస్ కింద ఫైల్ ను ఎలా సెలెక్ట్ చేసుకోవాలి, ఎలా ఫిల్టర్ చేయాలి, పేరును మార్చే అవకాశం, మార్చబడిన పేర్లను సరిచూసుకోవడం వంటి విషయాలు వినియోగదార్లకు మార్గదర్శనం చేస్తాయి.

పెద్దసంఖ్యలో ఫైళ్ల పేర్లను మార్చే అవసరమున్న ప్రతి పరిశోధకునికి ఈ క్విక్ ఫైల్ రీనేమర్ ఒక ముఖ్యమైన సాధనంగా ఉపయోగపడుతుంది. పేరు మార్చే సామర్థ్యాలు, వినియోగదారునికి స్నేహపూర్వక ఇంటర్‌ఫేస్‌ల కలయిక దీనిని ప్రొఫెషనల్, వ్యక్తిగత ఉపయోగాలు రెండింటికీ అవసరమైన అప్లికేషన్‌గా చెప్పవచ్చు. సంబంధిత లింకును ఇక్కడ పొందవచ్చు.

(https://medha.ciil.org/en/DesktopApplication/quickfilerenamer)

సాంద్ర: సాంద్ర ప్రసంగపాఠానికి సంబంధించింది. అంటే వినియోగదారుడు ఇచ్చిన ఏదైనా ఆడియో ఫోల్డర్‌లోని wav. ఫైల్‌లను MP3/AAC/WMA ఫైల్‌గా మారుస్తుంది. ఒకే ఫైల్ లేదా ఎక్కువ ఫైళ్లు ఉన్న ఫోల్డర్‌ను మార్చడానికి కూడా దీనిని ఉపయోగించవచ్చు. ఇచ్చిన మూల ఫైల్ ఫోల్డర్‌లోనే ఒకే ఫైల్గా మార్చిన ఫైల్ అందుబాటులోకి వస్తుంది. ఒక ఫోల్డర్‌ను ఇన్‌పుట్‌గా ఇచ్చినప్పుడు, కావలసిన అవుట్‌పుట్ ఫార్మాట్‌ను బట్టి, సోర్స్ ఫోల్డర్ పేరుతో '_mp3' లేదా '_AAC' లేదా '_WMA' పేరుతో సోర్స్ ఫోల్డర్ స్థానంలో కొత్త అవుట్‌పుట్ ఫోల్డర్ సృష్టించబడుతుంది. సాధ్యమైనంతగా ఫ్రాగ్మెంటేషన్‌ను నివారించడానికి అప్లికేషన్ డిఫాల్ట్‌గా సింగిల్ థ్రెడ్‌పై నడుస్తుంది. వేగంకోసం బ్యాచ్ ఫైల్‌లను రూపొందించుకుంటుంది. ఈ అప్లికేషన్ సహజ భాషా ప్రాసెసింగ్, స్పీచ్ టెక్నాలజీలపై పనిచేస్తున్న పరిశోధకులకు ఎంతగానో ఉపకరిస్తుంది. ఈ క్రింది లింకు ద్వారా దీనిని పొందవచ్చు. (https://medha.ciil.org/en/DesktopApplication/sandra)

తరంగ: తరంగ (wav. Metadata extractor) స్పీచ్ టెక్నాలజీకి సంబంధించిన అప్లికేషన్. ఇది ఒక wav. ఫైల్ కి సంబంధించి శీర్షిక దగ్గర నుండి నమూనా రేటు, బిట్ డెప్త్, ఛానెళ్ల సంఖ్య, ఆడియో ఆకృతి (ఫార్మట్)తో సహా ప్రతి వివరణాత్మక మెటాడేటాను సేకరిస్తుంది. ఇది ఫైలు పరిమాణం, వ్యవధి, బిట్ రేటు గురించి కూడా సమాచారాన్ని అందిస్తుంది. బహుళ సూచిక (మల్టిపుల్ డైరెక్టరీలు) లను పునరావృతంగా ప్రాసెస్ చేయగలదు. ఒకే సందర్భంలో అనేక ఫైళ్ళను విశ్లేషించగలదు. ప్రాసెస్ చేయబడిన అన్ని wav. ఫైళ్ళ వ్యవధినంతటినీ లెక్కించి చూపుతుంది. విశ్లేషించిన అన్ని wav. ఫైళ్ళ పరిమాణాన్నంతటినీ సంకలనం చేస్తుంది. ఆ నివేదికనంతటినీ ఒక స్ప్రెడ్ షీట్ కు పంపుతుంది. ఇది నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్, స్పీచ్ టెక్నాలజీలపై పనిచేస్తున్న పరిశోధకులు కార్పస్ పరిమాణం, వ్యవధిని తెలుసుకోవడానికి తరంగ అప్లికేషన్ ను ఉపయోగించవచ్చు. దీనిని ఈ క్రింది లింకు ద్వారా పొందవచ్చు.

(https://medha.ciil.org/en/DesktopApplication/taranga)

పదాన్వేషి: పదాన్వేషి అనేది కీవర్డ్ (సంకేతపదం) ను వెతకటానికి అభివృద్ధి చేయబడిన అప్లికేషన్. ఇది పాఠ్యదత్తాంశంలో కావలసిన పదాలను గుర్తించడానికి, సంగ్రహించడానికి రూపొందించబడింది. ఇది TXT లేదా XML ఫైళ్లను ప్రాసెస్ చేయగలదు. ఫైల్‌లు వేరే స్థానంలో ఉంటే కూడా వినియోగదారు బ్రౌజ్ చేయడంవల్ల దాని సరైన స్థానం తెలుస్తుంది. ఒక నిర్దిష్ట పదాన్ని కనుగొనడానికి వినియోగదారుడు కీవర్డ్ టెక్స్ట్ బాక్సులో పదాన్ని టైప్ చేసి సెర్చ్ బటన్‌పై క్లిక్ చేయాలి. ఆ పదాన్ని విడిగా కానీ, లేదా వివిధ విభక్తులతోపాటు కూడా శోధించడానికి అవకాశం కల్పిస్తుంది. సందర్భంతో పాటు కీవర్డ్ ప్రదర్శించే మొత్తం వాక్యం కనిపిస్తుంది. అవుట్‌పుట్ ను పదాన్వేషి అదేస్థానంలో కీవర్డ్ పేరు పెట్టబడిన టెక్స్ట్ ఫైల్‌గా సేవ్ చేసి ఇస్తుంది. దీనికి సంబంధించిన లింక్ ను ఈ క్రింద పొందవచ్చు.

(https://medha.ciil.org/en/DesktopApplication/padanveshi)

పదవృత్తి: పదవృత్తి అనేది పదపౌనఃపున్య గణాంకం (Word frequency counter). ఇది TXT, XML ఫైళ్ళ నుంచి మనం కోరిన పదాల సంబంధిత పౌనఃపున్యాలతో పాటు ప్రత్యేకమైన పదాల జాబితాను రూపొందిస్తుంది. ఇది పౌనఃపున్య జాబితా నుండి సంఖ్యలు, రోమన్ అక్షరాలు, విరామచిహ్నాలను మినహాయించి వినియోగదారునికి అందిస్తుంది. అవుట్పుట్ పదవృత్తి ప్రదేశంలో ట్యాబ్-వేరు చేసిన TXT ఫైల్ గా అందించబడుతుంది. అవుట్పుట్ ఫైల్లో వాటి పౌనఃపున్యాలతో పాటు విభిన్న టోకెన్ల జాబితా ఉంటుంది. అవుట్పుట్ ఫైలు పేరు 'పదవృత్తి' (విభిన్న టోకెన్లు ప్రాసెస్ చేయబడతాయి)గా చూపబడుతుంది. దీనికి సంబంధించిన లింకును ఈ క్రింద గమనించవచ్చు.

(https://medha.ciil.org/en/DesktopApplication/padavrtti)

5. మొబైల్ యాప్స్

ఎస్.డి.సి.పి (LDCIL-SDCP): ఎల్డీసిఐఎల్ రూపొందించిన తొలి మొబైల్ యాప్ ఎస్.డి.సి.పి (స్పీచ్ డేటా కలెక్షన్ పోర్టల్). ఇది భారతీయ భాషల్లో వాగ్భాషా నిధిని సేకరించటానికి రూపొందించిన వేదిక. ఇది గూగుల్ ప్లేస్టోర్ లో అందుబాటులో ఉంది. గతంలో స్పీచ్ డేటాను సేకరించటానికి మైక్రోఫోన్ ను చేతబట్టుకొని వ్యవహర్త ముందుంచి మాట్లాడించేవారు. ఈ ప్రక్రియలో నాణ్యమైన డేటా దొరుకుతుంది కానీ, సేకరణ చాలా ఖర్చుతోనూ, శ్రమతోనూ కూడివున్నది. అందువల్ల ఆ ప్రక్రియకు స్వస్తి పలుకుతూ వ్యవహర్తలకు, సేకర్తలకు సులభతరం కావటానికి ఈ యాప్ రూపొంద్చబడింది. దీని ద్వారా ప్రపంచంలో ఏ మూలనున్న వ్యవహర్తల నుంచైనా డేటా సేకరించవచ్చు.

ఈ యాప్ ను మొబైల్ లో దిగిమతి చేసుకున్న తర్వాత నిర్ణయించిన యూజర్ ఐడి, పాస్ వర్డ్ లను ఉపయోగించి లోపలికి ప్రవేశించాలి. లోపలి ముఖపేజీలో మై డాష్ బోర్డ్ క్రింద టాస్క్, కంప్లీట్, పెండింగ్ రికార్డ్ టాస్క్, పెండింగ్ రివ్యూ టాస్క్ వంటి సూచికలు కనిపిస్తాయి. పెండింగ్ రికార్డ్ టస్క్ పై నొక్కి, అసైన్డ్ చేసిన టాస్క్ విషయాన్ని తెరచి, రికార్డ్ బటన్ ఆన్ చేసుకొని నేరుగా విషయాన్ని చదవటమే. ఈ మధ్యలో ఏవైనా వ్యక్తిగత అసౌకర్యాలు కలిగితే రికార్డ్ బటన్ ను మళ్లీ నొక్కి కొంతసేపు నిలుపుదల తర్వాత మిగతా విషయాన్ని రికార్డు చేయవచ్చు. పూర్తిగా రికార్డు అయిన తర్వాత కింద ఉన్న సబ్మిట్ బటన్ ను నొక్కితే సరిపోతుంది. ఆ రికార్డు సేకర్తలకు చేరుతుంది. పెండింగ్ టాస్కులు, కంప్లీట్ టాస్కులు ముఖపేజీలో చూడవచ్చు. మనం రికార్డు చేసిన విషయాలను కూడా వినటానికి అవకాశం ఉంది. ఈ యాప్ ద్వారా సమయం, డబ్బు ఆదాతోపాటు నాణ్యమైన డేటా కూడా పొందవచ్చు.

6. ఉపసంహారం

భాష అనేది భావవ్యక్తీకరణ అనే భావనను దాటి సంస్కృతి, చరిత్ర, నాలెడ్జ్, ఐడెంటిటీ, ట్రాన్సులేషన్ టెక్నాలజీ డెవలెప్మెంట్, ఎఐ అనే స్థాయికి చేరింది. అయితే భారతదేశం లాంటి భాషా వైవిధ్యభరిత దేశంలో,

భాషల పరిరక్షణ, అభివృద్ధి, డిజిటలైజేషన్ వంటి బాధ్యతలు వహిస్తూ, భవిష్యత్ తరానికి భాషా సంపదను అందించే సందర్భంలో ఎల్డీసిఐఎల్ కీలకపాత్ర పోషిస్తోంది.

భాషాసాంకేతికకు సంబంధించి కార్పస్ ఇన్సైట్స్, కంపెన్డిమ్ ఆఫ్ ఎల్డీసిఐఎల్ సెంటెన్స్ అలైన్డ్ స్పీచ్ కార్పస్, కంపెన్డిమ్ ఆఫ్ లింగ్విస్టిక్ రిసౌర్సెస్ ఇన్ ఇండియన్ లాంగ్వేజెస్, లింగ్విస్టిక్ రిసౌర్సెస్ ఫర్ ఎఐ/ఎనెల్పి ఇన్ ఇండియన్ లాంగ్వేజెస్, కాస్ట్ అనాలసిస్ ఆఫ్ లింగ్విస్టిక్ రిసౌర్సెస్ వంటి భాషాసంబంధిత పుస్తకాలను, బుక్ చాప్టర్స్ ను, భారతీయ భాషల్లో పాఠ్య వాగ్దత్తాంశాలకు సంబంధించిన పుస్తకాలను ప్రచురించింది.

భాషాసాంకేతిక రంగానికి పునాదులైన టెక్స్ట్, స్పీచ్ డేటాసెట్స్ ను భారతీయ భాషలకు అందించటంతోపాటు, అవసరమైన సాంకేతికతను టూల్స్, అప్లికేషన్స్, యాప్స్ ను తయారుచేసి సమర్థవంతంగా నడుపుతున్నది. కేవలం టెక్స్ట్, స్పీచ్ కార్పొరాకాక, సెంటెన్స్ అలైన్డ్ స్పీచ్ కార్పొరా, పార్లల్ కార్పొరా, క్లాసికల్ లాంగ్వేజ్ కార్పొరా, పిఒయస్ టాగింగ్ కార్పొరా, ఛంకింగ్ కార్పొరా, సైన్ లాంగ్వేజ్ కార్పొరా వంటి వివిధ రకాల కార్పొరా మీద పనిచేస్తుంది.

స్పీచ్ డాటా, మోనోలింగ్వల్ టెక్స్ట్, కార్పస్, స్పీచ్ డేట్ అనోటేషన్, వేలిడేషన్, టిటిఎస్ వాయిస్ బిల్డింగ్, డిజిటైజేషన్, పార్ట్స్ ఆఫ్ స్పీచ్ అనోటేషన్, క్లాసికల్ లాంగ్వేజ్ కార్పస్ లను సృష్టిస్తున్నది. ఇక్కడ రూపొందిన అనువాదిక, అనులేఖిక, ఎస్డిసిపి, ఓసియార్, టెక్స్ట్ నార్మలైజర్, కార్పస్ బ్రౌజర్స్, కోబోర్డ్ లే-అవుట్స్ వంటి ఇంకా అనేక ఇతర భాషాటూల్స్ యూజర్-ఫ్రెండ్లీగా రూపుదిద్దుకుంటూ, ప్రభుత్వ ప్రాజెక్టులకు, పరిశోధకులకు, విద్యార్థులకు, పరిశ్రమలకు విలువైన సహకారాన్ని అందిస్తున్నాయి. ఇవి కేవలం తాత్కాలిక పరిష్కారాలు మాత్రమే కాకుండా, భవిష్యత్తు నేచురల్ లాంగ్వేజ్ ప్రాసెస్, మెషిన్ ట్రాన్సులేషన్స్, స్పీచ్ రికగ్నిషన్స్, లాంగ్వేజ్ లెర్నింగ్ వంటి రంగాల్లో బలమైన పునాదులుగా నిలుస్తున్నాయి.

భారత ప్రభుత్వం యొక్క 'డిజిటల్ భారత్' లక్ష్యాన్ని నెరవేర్చడంలో, ప్రతి పౌరుడు తమ మాతృభాషలో సమాచారాన్ని పొందేలా చేసే మహత్తర ప్రయత్నానికి ఎల్డీసిఐఎల్ చేసిన కృషి వెనుకబడిన భాషలకు కొత్త ఊపిరినివ్వడమే కాక, భారతీయ భాషల గౌరవాన్ని ప్రపంచపటంలో నిలిపే లక్ష్యం. అందుకు ఇక్కడ జరుగుతున్న మదర్ టంగ్ పార్లల్ కార్పస్ ట్రాన్సులేషన్ వంటి ప్రాజెక్టులు సాక్ష్యాలు. ఈ సందర్భంగా భాషాభివృద్ధి అనేది కేవలం ఒక సంస్థకే పరిమితం కాదని, ఇది ప్రభుత్వాలు, విద్యాసంస్థలు, పరిశోధకులు, భాషాభిమానుల అందరి సమిష్టి బాధ్యత అని గుర్తించాలి.

ఎల్డీసిఐఎల్ ఈ బాధ్యతను ముందుండి నడిపిస్తూ, భవిష్యత్ తరం అవసరాలకు అనుగుణంగా భాషా సాంకేతికతను సమర్థవంతంగా అందించడంలో మార్గదర్శకంగా నిలుస్తుంది.

7. పాదసూచికలు

1. https://www.ldcil.org/english/aboutUs.aspx
2. https://www.ldcil.org/english/workInProgress.aspx

8. ఉపయుక్తగ్రంథసూచి

1. రాధాకృష్ణ బూదరాజు, ఆధునిక వ్యవహారకోశం (ఇంగ్లీషు-తెలుగు), ప్రాచీ పబ్లికేషన్స్, 2008
2. LDC-IL Corpus Insights, Narayan Kumar Choudhary, CIIL, 2025, Mysore
3. Linguistic Resources for AI/NLP in Indian Languages, Narayan Kumar Choudhary, CIIL, 2019, Mysore
4. The Mother Tongue Parallel Text Corpus of India Vol. I, Narayan Kumar Choudhary, CIIL, 2025, Mysore
5. Dr. Modugu Kasimbabu, Rajesha R, Manasa G, Narayan Choudhary, Prof. Shailendra Mohan. 2025. A Gold Standard Telugu Raw Text Corpus Vol.II, Central Institute of Indian Languages, Mysore
6. https://ldcil.org/

***

గమనిక: ఈ పత్రికలోని వ్యాసాలలో అభిప్రాయాలు రచయితల వ్యక్తిగతమైనవి.

వాటికి సంపాదకులు గానీ, పబ్లిషర్స్ గానీ ఎలాంటి బాధ్యత వహించరు.