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-04 | April 2025 | Peer-Reviewed | ISSN: 2583-4797

Article Link: https://www.auchithyam.com/2025/apr_2025/full_papers/april25_03.php

DOI: 10.5281/zenodo.15202242

3. అంతర్జాలంలో తెలుగు భాషానిధులు, సాంకేతికవనరులు

డా. మోదుగు కాశీంబాబు

తెలుగు రిసోర్స్ పర్సన్, ఎల్డీసిఐయల్ (LDCIL),
భారతీయ భాషాసంస్థ (CIIL),
మైసూర్, కర్నాటక.

సెల్: +91 9908683093. Email: kasimtelugu@gmail.com

సమర్పణ (D.O.S): 20.03.2025 ఎంపిక (D.O.A): 31.03.2025 ప్రచురణ (D.O.P): 01.04.2025

వ్యాససంగ్రహం:

భాషకు సాంకేతికతను అందించే సందర్భంలో దాదాపు నలభై సంవత్సరాల నుంచి భాషాసాంకేతిక అధ్యయనాలు జరుగుతున్నాయి. ఇందులో తెలుగు పాఠ్య, వాగ్భాషా నిధుల్ని సమకూర్చటం ప్రధాన విషయమైంది. ఈ క్రమంలో అనేక ప్రభుత్వ, ప్రైవేట్ సంస్థలు పూనుకొని తెలుగుభాషాభివృద్ధి కోసం అనేక పాఠ్య, వాక్ డేటాసెట్లను తయారుచేశాయి. ఈ డేటాసెట్లు ప్రధానంగా ఏఐ వంటి అత్యాధునిక భాషామోడళ్ళకు అత్యంత ఆవశ్యకం. భారతీయసంస్థలే కాక విదేశీసంస్థలు కూడా తెలుగుభాషాభివృద్ధి వరుసలో ఉండటం గమనార్హం. ఈ సంస్థలలో జరిగిన భాషానిధులను వివరించటం ఇందులోని ప్రధానాంశం. ఇక్కడ మరో అంశం సాంకేతిక వనరులు. తెలుగుకు ప్రస్తుతం మిగతా భారతీయ భాషలతో పోల్చితే మెరుగైన వనరులే అందుబాటులో ఉన్నాయి. వందల్లో నిఘంటువులు, లక్షల్లో ముద్రితగ్రంథాలు, పత్రికలు, వ్యాసాలు ఉచితంగా అందుబాటులో ఉన్నాయి. ఇవన్నీ తెలుగు పరిశోధనను మరింత అభివృద్ధి చేయటానికి ఉపకరిస్తాయి. నిజానికి ఈ వ్యాసం అంతర్జాలంలో అందుబాటులో ఉన్న ప్రధాన జాలవేదికల ఆధారంగా రూపొందించబడింది. తెలుగులో ఈ విషయంపై వచ్చిన వ్యాసాలు దాదాపుగా అందుబాటులో లేవు. ఇందులో ప్రధానంగా భాషిణి, ఏఐ4భారత్, ఐఐటీ మద్రాస్, ఐఐటీ హైదరాబాద్, ఎల్డీసిఐయల్, ఐయల్ సిఐ వంటి ప్రభుత్వ సంస్థలు చేసిన కృషిని, వాణి, తెలుగు ఎల్.ఎల్.ఎం ల్యాబ్స్ వంటి స్వచ్ఛందసంస్థలు, ఆస్కార్, డాటా ఓషన్, ఫ్యూచర్ జీ ఏఐ, లిష్జిగ్ వంటి విదేశీ సంస్థల కృషిని చూడవచ్చు. ఇందులోని వివరాల ఆధారంగా అంతర్జాలంలో ఉన్న తెలుగు భాషానిధులను, సాంకేతిక వనరులను పరిశోధకులు వినియోగించుకునే అవకాశం ఉంటుంది. తెలుగు పై కార్పస్ సంబంధిత పరిశోధనలకు ఇది ఆధారమవుతుంది. తెలుగు భాషాభిమానులకు, పండితులకు, పరిశోధకులకు భాషాసాంకేతికత పై అభిరుచిని, ఆసక్తిని కలిగించటానికి ఈ వ్యాసం ఉపకరిస్తుందని భావిస్తున్నాను.

Keywords: కార్పస్-భాషానిధి; టెక్నికల్ రిసౌర్సెస్- సాంకేతికవనరులు; ఏఐ ఆర్టిఫిషియల్ ఇంటిలిజెన్స్-కృత్రిమమేధ; టూల్స్-సాధనాలు, ఉపకరణాలు; డేటాసెట్స్-దత్తాంశసమితులు; టెక్స్ట్-పాఠం, పాఠ్యం; స్పీచ్-వాక్కు

1. ప్రవేశిక:

భాష అనేది కేవలం సమాచార మాధ్యమమే కాదు, ఒక జాతి చరిత్ర, జ్ఞానం (Knowledge), సాంస్కృతికవారసత్వం, ఆ జాతి సమస్త అస్తిత్వానికి సాక్ష్యం. భాషవల్లనే జాతి అస్తిత్వం నిలబడుతుంది. ప్రస్తుతం ఆ భాషకు సాంకేతికత తోడై ఆధునికసమాజంలో ఎన్నో విప్లవాత్మక మార్పుల్ని సృష్టిస్తోంది. సాంకేతికత అందుబాటులో లేనికారణంగానే గతంలో ఈ ప్రపంచం ఎన్నోభాషల్ని, జాతుల్ని కోల్పోయింది. సాంకేతికత ప్రపంచంలోని అనేక భాషల్ని ఒకటిచేస్తూ, మరింత సౌలభ్యంగా, శాస్త్రీయంగా తీర్చిదిద్దుతూ, నూతన ఆవిష్కరణలకు నాంది పలుకుతుంది. ఈక్రమంలోనే అనేక ప్రభుత్వ, ప్రైవేటురంగ సంస్థలు భాషానిధుల్ని సృష్టించే వనరులకేంద్రాలుగా వెలశాయి. ఆ సంస్థల ప్రణాళికలు, అవి సమీకరించిన భాషానిధులను, ప్రపంచభాషల అభివృద్ధి కోసం వచ్చిన సాంకేతిక సాధనాలను వివరించడం ఉద్దేశం. సమర్థవంతమైన భాషా అధ్యయనానికి, విశ్లేషణకు భాషా వనరులు అవసరం. మరింత నాణ్యమైన సామగ్రిని (డేటా) అభివృద్ధి చేయడానికి తగిన సమయం, కృషి కూడా అవసరమే. లేఖనపద్ధతి (రైటింగ్ సిస్టమ్స్), ముద్రణయంత్రం (ప్రింటింగ్ ప్రెస్), కంప్యూటర్లు వంటివి వచ్చిన తర్వాత భాషా సమాచారాన్ని నిక్షిప్తంచేయడం మరింత సులభమైంది. ప్రామాణిక భాషా సామగ్రి, సాధనాల (టూల్స్) సమగ్ర జాబితా భాషా అధ్యయనాలను ముందుకు తీసుకెళ్లడానికి ప్రయోజనకరంగా ఉంటుంది. ఒక నిర్దిష్ట భాష పరిణామాన్ని, పెరుగుదలను అర్థం చేసుకోవడానికి, పరిశోధనను విస్తరించడానికి, భాషాసాంకేతికతను మెరుగుపరచడంలో ఈ వనరులు ఎంతో కీలకమైనవని చెప్పవచ్చు.

భాషా వనరులు అనేవి లిఖితపూర్వకంగానూ, వాగ్రూపాల్లోనూ మానవభాషలను నడిపించడానికి, విశ్లేషించడానికి, ఉత్పత్తి చేయడానికి ఉపయోగించే ముఖ్యమైన డేటాసెట్‌లు, సాధనాలు, సాంకేతికతలు. అవి సహజ భాషా ప్రాసెసింగ్ (NLP), కంప్యూటేషనల్ లింగ్విస్టిక్స్, ఆర్టిఫిషియల్ ఇంటెలిజెన్స్ (AI) మరియు మెషిన్ లెర్నింగ్ (ML) అప్లికేషన్‌లకు పునాదిగా పనిచేస్తాయి. యంత్రానువాదం, స్పీచ్ రికగ్నిషన్, టెక్స్ట్-టు-స్పీచ్ (TTS), సెంటిమెంట్ విశ్లేషణ వంటి వివిధ భాషా ఆధారిత సాంకేతికతలను అభివృద్ధి చేయడంలో ఈ వనరులు కీలకమైనవి.

2. వనరుల అభవృద్ధి సవాళ్లు:

భాషా అవరోధాలను అధిగమించడానికి భాషా సాంకేతికత లాభదాయకమైన, సామాజికంగా ప్రయోజనకరమైన పరిష్కారాన్ని అందిస్తుంది. తెలుగు వంటి సంక్లిష్ట భాషలకు ఖచ్చితమైన భాషా నమూనాలను అభివృద్ధి చేయడంలో భాషా డేటా డిజిటలైజేషన్ సవాళ్లను ఎదుర్కొంటుంది. కంప్యూటర్లు మానవభాషను అర్థం చేసుకోవడానికి వీలు కల్పించడంలో మొదటి దశ ఎన్కోడింగ్, ఇది తెలుగు అక్షరాలకు ప్రత్యేక కోడ్లను కేటాయించే యూనికోడ్ పథకం ద్వారా సాధ్యమవుతుంది.

3. భాషావనరులు:

తెలుగు డేటాను విశ్లేషించడానికి, సాంకేతిక పరిజ్ఞానాన్ని అభివృద్ధి చేయడానికి ఎంతో సామగ్రి (న్యూమరల్ మెటీరియల్స్) ఉంది. ఈ వనరులలో పాఠ్యం (టెక్స్ట్), వాక్కు (స్పీచ్) భాషానిధి, నిఘంటువులు, ఆంటాలజీలు, మల్టీమీడియా డేటాబేస్ల సేకరణకు, తయారీకి, విశ్లేషణకు సాఫ్ట్వేర్లు ఉన్నాయి. రియల్ టైమ్ లాంగ్వేజ్ వినియోగానికి ప్రాతినిధ్యం వహించే లింగ్విస్టిక్ కార్పస్, వివిధ భాషా సాంకేతిక పరిజ్ఞానాలను అభివృద్ధి చేయడానికి కీలకమైనది.

భాషా సాంకేతికత యొక్క ప్రధాన అనువర్తన ప్రాంతాలలో స్పెల్, గ్రామర్ చెకింగ్, స్పీచ్ రికగ్నిషన్, సింథసిస్, యంత్రానువాదం, సమాచార పునరుద్ధరణ ఉన్నాయి. మానవులు, కంప్యూటర్ల మధ్య కమ్యూనికేషన్, పరస్పర చర్యను పెంచే ఈ సాధనాలకు భాషా వనరులే పునాది.

4. ప్రముఖ భాషావనరుల సంస్థలు – వాటి ప్రణాళికలు:

భాషానిధి (కార్పస్) అనేది భాషా విశ్లేషణ, సహజ భాషా విశ్లేషణ (ఎన్ఎల్పి) కోసం ఉపయోగించే గ్రంథాల నిర్మాణాత్మక సేకరణ. ఇందులో ఒకేభాషకు చెందిన గ్రంథాలు, బహుభాషల (మల్టీలింగ్వల్) గ్రంథాలు ఒకేగ్రంథంలో అనేక భాషలతోకూడినవి, వివరణాత్మక గ్రంథాలు, సాహిత్యంతోపాటు వైద్యం, రాజకీయం, ఆర్థికం వివిధ శాస్త్రాలకు సంబంధించిన గ్రంథాలు, ప్రాచీనభాషను అందించే గ్రంథాలను ఒకచోట ఉంచుతారు. భాషా సంబంధిత విషయాలను అధ్యయనం చేయడం, యంత్ర అభ్యాస నమూనాలకు శిక్షణ ఇవ్వడం, భాషా నమూనాలను నిర్మించడం, ఎన్ఎల్పి సాధనాలను (NLP Tools) అభివృద్ధి చేయడం వంటి పనులకు ఈ భాషానిధి ఉపయోగించబడుతుంది. ప్రపంచవ్యాప్తంగా వివిధ ప్రభుత్వ, ప్రైవేట్ భాషాసంస్థలు భారతీయభాషలకు పాఠ్యవాగ్భాషానిధుల్ని తయారుచేసాయి. తెలుగుకు సంబంధించి కేవలం భారతీయసంస్థలేకాకుండా విదేశీసంస్థలు కూడా ఉన్నాయి. వాటిని ఈ కింద గమనించవచ్చు.

5. తెలుగు భాషానిధులు:

తెలుగులో పాఠ్య, వాగ్భాషానిధుల సేకరణ, సహజ భాషా విశ్లేషణ అనేది సుమారు నలభై ఏళ్లనుండి జరుగుతున్నది. ఈ రంగంలో ప్రస్తుతం ప్రభుత్వరంగ సంస్థల కృషి అభివృద్ధి చెందుతున్నది. దేశీయ విదేశీయ సంస్థలలో తెలుగులో అందుబాటులో ఉన్న భాషానిధుల్ని, వివిధ భాషాసంస్థల్ని పరిశీలిద్దాం.

భాషిణి (BHASHINI)

'భాషిణి' భారతీయ భాషల అనువాద వేదిక. ఇది భారతదేశ భాషా ఇంటర్ఫేస్ సంక్షిప్తరూపం. భారతదేశంలోని 22 షెడ్యూల్డ్ భాషల్లో భాషా అవరోధాలను విచ్ఛిన్నం చేసి, వివిధ భాషలు మాట్లాడేవారి మధ్య సంభాషణలను అందిస్తున్న కృత్రిమమేధ ఆధారిత భాషా సాంకేతిక అనువాద వ్యవస్థ. ఇది ఆండ్రాయిడ్, ఐఓఎస్ యాప్ ల ద్వారా అందుబాటులో ఉంది. నేషనల్ లాంగ్వేజ్ టెక్నాలజీ మిషన్ కింద 2022 జూలైలో గౌరవ

ప్రధానమంత్రి శ్రీ నరేంద్రమోదీ చేతుల మీదుగా ప్రారంభమైంది. తెలుగుకు 64,08,913 వాక్యాల పాఠ్యభాషానిధిని, 3,053.676 వాల్యూ కలిగిన ASR డేటాసెట్లను, 850.116 వాల్యూ కలిగిన ASR అన్ లేబుల్డ్ డేటాసెట్లను కలిగివుంది.

ఎఐ4భారత్ (AI4Bharat)

ఐఐటీ మద్రాస్ లోని రీసెర్చ్ ల్యాబ్ ఏఐ4భారత్ ఓపెన్ సోర్స్ కార్యక్రమాల ద్వారా భారతీయ భాషలకు ఏఐ టెక్నాలజీని అభివృద్ధి చేయడానికి స్థాపించబడింది. ఈ ల్యాబ్ విస్తృతమైన డేటాసెట్లు, టూల్స్, అత్యాధునిక నమూనాలను అభివృద్ధి చేసి విడుదల చేసింది. అనువాదం, సహజ భాషా అవగాహన మరియు జనరేషన్, అనువాదం, ఆటోమేటిక్ స్పీచ్ రికగ్నిషన్, స్పీచ్ సింథసిస్ వంటి విషయాలపై కృషిచేస్తూ ప్రపంచవ్యాప్తంగా గుర్తింపు పొందింది. ఉన్నతస్థాయి సదస్సులతో విద్యారంగం, వివిధ ప్రభుత్వ రంగాలలో గణనీయమైన ప్రభావాన్ని చూపింది. తెలుగులో 16,278.7 మిలియన్ పదాల పాఠ్యభాషానిధిని, 136 గంటల వాగ్భాషానిధిని రూపొందించింది.

సి-డాక్ హైదరాబాద్ (C-DAC)

సెంటర్ ఫర్ డెవలప్మెంట్ ఆఫ్ అడ్వాన్స్డ్ కంప్యూటింగ్ (సి-డాక్) అనేది ఎలక్ట్రానిక్స్ అండ్ ఇన్ఫర్మేషన్ టెక్నాలజీ మంత్రిత్వశాఖ ఆధ్వర్యంలో పనిచేసే ఒక భారతీయ స్వయంప్రతిపత్తి కలిగిన సంస్థ. సి-డాక్ 1987 నవంబరులో రూపొందించబడింది. ఇది అమెరికా తర్వాత ప్రపంచంలోనే అత్యంత శక్తిమంతమైన సూపర్ కంప్యూటర్ భారత్ కు ఉందని నిరూపించింది. సి-డాక్ అడ్వాన్సుడ్ కంప్యూటింగ్, సాఫ్ట్ వేర్ డెవలప్మెంట్ రంగంలో పలు కోర్సులను అందిస్తోంది. పవర్ సిస్టమ్స్, ఎంబెడెడ్ సాఫ్ట్ వేర్, ఓపెన్సోర్స్ ఆపరేటింగ్ సిస్టమ్స్, హై పెర్ఫార్మెన్స్ కంప్యూటింగ్, సెమీకండక్టర్ డిజైన్ అండ్ డెవలప్మెంట్ రంగాల్లో సి-డాక్ సృష్టించిన అత్యాధునిక సాంకేతిక పరిజ్ఞానాల వాణిజ్యీకరణకు ఇది ఒక ప్రోగ్రామ్ ను ఏర్పాటు చేస్తుంది. తెలుగులో 8.22 MB వాగ్భాషానిధిని రూపొందించింది.

ఒపస్ (Opus):

OPUS అనేది జాలవేదిక నుండి అనువదించబడిన గ్రంథాల సేకరణకు సంబంధించిన ప్రాజెక్టు. ఇది ఉచిత ఆన్ లైన్ డేటాను మార్చగలదు, సేకరించగలదు. అలాగే లభ్యమవుతున్న సమాంతర కార్పస్ ను పౌరసమాజానికి అందించగలదు. OPUS బహిరంగ మూల ఉత్పత్తులపై ఆధారపడి ఉంటుంది. ఇది పాఠ్యనిధిని బహిరంగ విషయాల టోకుగా కూడా దిగుమతి చేస్తుంది. ఒపస్ లో ఆంగ్లం నుంచి తెలుగుకు సమాంతరంగా అనువదించిన 54,363,342 వాక్యాలు అందుబాటులో ఉన్నాయి.

వాణి (VAANI):

వాణి అనేది సమ్మిళిత డిజిటల్ ఇండియా కోసం భాషాకృత్రిమమేధ సాంకేతికపరిజ్ఞానాన్ని, కంటెంట్ ను ముందుకు తీసుకెళ్లడానికి ఇండియన్ ఇన్స్టిట్యూట్ ఆఫ్ సైన్స్ (IISc), బెంగళూరు మరియు ఏఐ$రోబోటిక్స్ టెక్నాలజీ పార్క్ (ARTPARK) సంయుక్తంగా నిర్వహిస్తున్న ప్రాజెక్ట్. ఇది భారతదేశంలో అందుబాటులో ఉన్న భాషల నిజమైన వైవిధ్యాన్ని సంగ్రహిస్తోంది. 150,000 గంటలకు పైగా వాగ్భాషానిధిని సృష్టించాలన్నది దీని లక్ష్యం. ప్రస్తుతం దీనిలో కొంతభాగం స్థానిక లిపులలో అనువదించబడుతుంది. ఇది భాష, విద్య, పట్టణ-గ్రామీణ, వయస్సు, లింగ వైవిధ్యాన్ని నిర్ధారిస్తుంది. భారతదేశంలోని మొత్తం 773 జిల్లాల్లోని సుమారు ఒక మిలియన్ ప్రజల నుండి వాగ్భాషానిధి, పాఠ్యనిధిని సేకరించారు. ఈ డేటాసెట్లు ఓపెన్ సోర్స్ చేయబడతాయి. డేటా ప్రస్తుత వెర్షన్ ఇక్కడ ఓపెన్ సోర్స్ చేయబడింది. రాబోయే రోజుల్లో భాషిణి (జాతీయ భాషా అనువాద సంస్థ, MeiTY) వంటి వేదికల ద్వారా అందుబాటులో ఉంచాలని ఆశిస్తున్నది. తెలుగు రాష్ట్రాలతోపాటు దేశంలోని ఇతర రాష్ట్రాలలో తెలుగు మాట్లాడేవారినుంచి 1,487.08 గంటల వాగ్భాషానిధిని సేకరించి అందుబాటులో ఉంచారు.

కెగిల్ (KAGGLE)

అమెరికాకు చెందిన ఆంటోని గోల్డ్ బ్లూమ్ 2010లో కెగిల్ సంస్థను స్థాపించాడు. ఇది గూగుల్ ఎల్.ఎల్.సి కింద పనిచేస్తుంది. డేటా సైన్స్, మెషిన్ లెర్నింగ్ రంగాల్లో పోటీకి, పరస్పర సహకారానికి, నూతనావిష్కరణలు పంచుకోవడానికి ఏర్పాటుచేయబడిన అంతర్జాల వేదిక. ఇక్కడ వినియోగదారులు డేటాసెట్లను పొందవచ్చు, ప్రచురించవచ్చు అలాగే వెబ్-ఆధారిత డేటా సైన్స్ రంగంలో నమూనాలను అన్వేషించవచ్చు, నిర్మించవచ్చు. ఇంకా ఇతర డేటా శాస్త్రవేత్తలు, మెషిన్ లెర్నింగ్ ఇంజనీర్లతో కూడా కలిసి పనిచేయవచ్చు. ఇది డేటా సైన్స్ సవాళ్లను పరిష్కరించడానికి అనుమతినిస్తుంది. ఇక్కడ 117 తెలుగు వాగ్భాషానిధికి సంబంధించిన డేటాసెట్లు అందుబాటులో ఉన్నాయి.

స్కెచ్ ఇంజిన్ (Sketch Engine)

స్కెచ్ ఇంజిన్ 2003 లో ప్రముఖ భాషాశాస్త్ర పరిశోధకుడు, అనువాదకుడు అయిన ఆడమ్ కిల్గారిఫ్ స్థాపించాడు. ఇది లెక్సికల్ కంప్యూటింగ్ చేత అభివృద్ధి చేయబడిన కార్పస్ మేనేజర్, టెక్స్ట్ అనాలిసిస్ సాఫ్ట్ వేర్. దీని అల్గోరిథంలు బిలియన్ల పదాల (టెక్స్ట్ కార్పోరా) ప్రామాణిక గ్రంథాలను విశ్లేషిస్తాయి. భాషలో ఏది విలక్షణమైనది, ఏది అరుదైనది, ఏది అసాధారణమైనది, ఉపయోగకరమైనది అని తక్షణమే గుర్తిస్తుంది. ఇది పాఠ్య విశ్లేషణ, టెక్స్ట్ మైనింగ్ అనువర్తనాల కోసం కూడా రూపొందించబడింది. ప్రపంచవ్యాప్తంగా ఉన్న ప్రచురణకర్తలు, విశ్వవిద్యాలయాలు, అనువాద సంస్థలు, జాతీయ భాషా సంస్థలకు ఇది మొదటి ఎంపికగా ఉంది. వందకుపైగా భాషలలో 800 రెడీ-టు-యూజ్ కార్పొరాలో ఒక ట్రిలియన్ పదాలను కలిగి ఉంది. ఇక్కడ 12,68,07,158 పదాల తెలుగు పాఠ్యభాషానిధి అందుబాటులో ఉంది.

ఫ్యూచర్ బీ ఎఐ (FUTUREBEE AI)

ఏఐ డేటా సోర్సింగ్, లోకలైజేషన్ రంగంలో ఫ్యూచర్ బీఏఐ ప్రముఖ సేవలను అందించే సంస్థ. దీని స్థాపకుడు జెసల్ టక్కర్. ఇది అతిపెద్ద OTS ట్రైనింగ్ AI డేటాసెట్ లు, కస్టమ్ ట్రైనింగ్ డేటా సేకరణను కలిగివుంది. వివిధ ఎనొటేషన్ సేవలతో AI స్టార్టప్ లు, సంస్థలకు సేవలు అందిస్తున్నది. ఫ్యూచర్ బీ ఏఐ మానవ కేంద్రీకృత కృత్రిమ మేధస్సు ఆధారిత ప్రపంచాభివృద్ధికి దోహదం చేస్తున్నది. పలు ఏఐ ఆధారిత సంస్థలతో కలిసి పనిచేయడానికి సిద్ధంగా ఉంటుంది. AI అభివృద్ధిలో భాగంగా రెండువేలకు పైగా డేటాసెట్ లను అందిస్తున్నది. ఆంగ్లానికి సమాంతరంగా తెలుగులో 6 లక్షల పదాల పాఠ్యభాషానిధిని, 300 పైగా తెలుగు వాగ్భాషానిధిని కలిగివున్నది.

డాటా ఓషన్ ఎఐ (DATAOCEAN AI)

ఆటోమేటిక్ స్పీచ్ రికగ్నిషన్ (ASR), టెక్స్ట్-టు-స్పీచ్ (TTS), నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్ (NLP) వంటి వివిధ మెషిన్ లెర్నింగ్ పనుల కోసం రూపొందించిన డేటాసెట్ల సమాహారమే డేటాఓసియన్ఏఐ కార్పస్. ఈ డేటాసెట్లు విస్తృతస్థాయి భాషలను, యాసలను, దృశ్యాలను కవర్ చేస్తాయి. స్మార్ట్ హోమ్ పరికరాలు, స్వతంత్ర డ్రైవింగ్, మరెన్నో అనువర్తనాల కోసం ఎఐ నమూనాలను అభివృద్ధి చేయడానికి ఇవి ఉపయోగపడతాయి. ఇది తెలుగు కన్వర్షనల్ స్పీచ్ రికగ్నిషన్ కార్పస్ (డెస్కుటాప్) 425 గంటలు; తెలుగు కన్వర్షనల్ స్పీచ్ రికగ్నిషన్ కార్పస్ (టెలిఫోన్) 100 గంటలు; తెలుగు స్పీచ్ రికగ్నిషన్ కార్పస్ (డెస్కుటాప్) 118 గంటలు; తెలుగు స్పీచ్ రికగ్నిషన్ కార్పస్ (మొబైల్) 118 గంటలు కలిగి ఉంది.

ఇండిక్ ఎయస్ఆర్ (INDICASR)

ఇండిక్ ఏఎస్ ఆర్ అనేది భారతీయ భాషల కోసం ప్రత్యేకంగా రూపొందించిన స్పీచ్ రికగ్నిషన్ సిస్టమ్. భారతీయ భాషల కోసం సమగ్రమైన, సమ్మిళిత బహుభాషా ప్రసంగ డేటాసెట్ ను రూపొందించడానికి ఉద్దేశించిన ఇండిక్ వాయిస్ ప్రాజెక్టులో భాగమిది. భారత రాజ్యాంగం 8వ షెడ్యూలులో పేర్కొన్న 22 భాషలకు ఇండిక్ ఏఎస్ ఆర్ సహకారాన్నందిస్తుంది. దీనికి భారత ప్రభుత్వ ఎలక్ట్రానిక్స్, ఇన్ఫర్మేషన్ టెక్నాలజీ మంత్రిత్వశాఖ నిధులను అందిస్తుంది. ఎక్ స్టెప్ ఫౌండేషన్, నీలేకని ఫిలాంత్రోపీస్ లు దీనికి సహకారాన్నందిస్తున్నాయి. ఈ డేటాసెట్ షెడ్యూల్డ్ భాషలలోని వేలాది మంది వక్తల నుండి సహజమైన, స్వతస్సిద్ధమైన ప్రసంగాలను కలిగి ఉంది. యూట్యూబ్ లో లభిస్తున్న వివిధ తెలుగు ఇంటర్వ్యూ ప్లేలిస్టుల నుండి తీసుకున్న 94 గంటల వాగ్భాషానిధి అందుబాటులో ఉంది.

ఎల్రా (ELRA)

ELRA ఒక లాభాపేక్షలేని భాషా వనరుల సంస్థ. ఇది 1995లో స్థాపించబడింది. హ్యూమన్ లాంగ్వేజ్ టెక్నాలజీస్ (HLT) కోసం భాషావనరుల (LRS) కమ్యూనిటీకి అందుబాటులో ఉంచడం దీని ప్రధాన

లక్ష్యం. అందువల్ల ఇది భాషావనరులకు సంబంధించి అనేక రకాల కార్యకలాపాలను నిర్వహిస్తుంది. అందులో గుర్తించడం & పంపిణీచేయడం (ఐడెంటిఫికేషన్ & డిస్ట్రిబ్యూషన్), ఉత్పత్తి & సక్రమత (ప్రొడక్షన్ & వాలిడేషన్), సాంకేతిక మూల్యాంకనం (టెక్నాలజీ ఎవాల్యుయేషన్), HLT పై సమాచార వ్యాప్తి వంటివి ఉన్నాయి. ఇక్కడ 118 గంటల తెలుగు వాగ్భాషానిధి అందుబాటులో ఉంది.

తెలుగు-ఎల్ఎల్ఎం-ల్యాబ్స్ (Telugu-LLM-Labs)

రవితేజ, రాంశ్రీ గౌతమ్ గొల్ల స్థాపించిన ఒక స్వతంత్ర సహకార సంస్థ తెలుగు-LLM-ల్యాబ్స్. ఇది తెలుగు భాష కోసం సహజ భాషా ప్రాసెసింగ్ (NLP)ని అభివృద్ధి చేయడమే లక్ష్యంగా పెట్టుకుంది. ప్రపంచవ్యాప్తంగా తెలుగు మాట్లాడేవారి కోసం AI అప్లికేషన్‌లను మెరుగుపరచడానికి ఓపెన్-సోర్స్ డేటాసెట్‌లు, నమూనాలను సృష్టించడం దీని ప్రాథమిక లక్ష్యం.

లింగ్విస్టిక్ డేటా కన్సార్షియం ఫర్ ఇండియన్ లాంగ్వేజెస్ (LDC-IL)

ఎల్డీసీ-ఐఎల్ అనేది భారతప్రభుత్వం ఉన్నతవిద్యాశాఖకు చెందిన ఒక ప్రాజెక్టు. ఇది భారతీయ భాషాసంస్థ మైసూరులో ఉన్నది. ఈ ప్రాజెక్టు భారతీయ భాషలలో వివిధ శాస్త్రరంగాలనుంచి నాణ్యమైన పాఠ్య వాగ్భాషానిధుల్ని తయారుచేసి అభివృద్ధి చేసింది. భాషాసాంకేతికతకు అవసరమైన సామగ్రిని అందించటంతో పాటు, వివిధ సాంకేతిక టూల్సును తయారుచేసి పరిశోధకులకు, పండితులకు అందుబాటులో ఉంచుతుంది. 2019 ఏప్రిల్ 4 నుంచి తన డేటా డిస్ట్రిబ్యూషన్ పోర్టల్ ద్వారా కృత్రిమ మేధస్సు (AI), నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్ (NLP) కోసం భాషా వనరులను ప్రధానంగా భారతీయ భాషల్లో అందుబాటులో ఉంచడం ప్రారంభించింది. తెలుగులో సాహిత్యం, వాణిజ్యం, సమాచారం, అధికారిక పత్రాలు, సాంకేతిక శాస్త్రవిజ్ఞానం వంటి రంగాలనుంచి 60,24,604 పదాల సహజ పాఠ్యభాషానిధి, 22:44 నిమిషాల సహజ వాగ్భాషానిధి అందుబాటులో ఉంది.[1]

ఇండియన్ లాంగ్వేజెస్ కార్పొర ఇనిషియేటివ్ (ILCI)

టిడిఐఎల్ ప్రారంభించిన ఇండియన్ లాంగ్వేజెస్ కార్పొరా ఇనిషియేటివ్ (ఐఎల్సిఐ) ప్రామాణిక ఫ్రేమ్ వర్కుల ఆధారంగా జాతీయ పాఠ్యభాషానిధిని సృష్టించింది. ఈ పని రెండు విడతలుగా జరిగింది. మొదటి విడతలో ఇంగ్లీష్ తో సహా 12 ప్రధాన భారతీయ భాషలలో సమాంతర వివరణాత్మక భాషానిధిని (Parallel Annotated Corpora) అభివృద్ధి చేసింది. భాషాభాగాలను (పార్ట్-ఆఫ్-స్పీచ్ (పిఓఎస్) వివరించటం కోసం భారతదేశ జాతీయ ప్రమాణాలను (BOS) ఉపయోగించింది. రెండవ విడతలో సుమారు 27 మిలియన్ల సమాంతర సంక్షిప్త పదాల పాఠ్యభాషానిధిని రూపొందించింది. ఈ ప్రాజెక్టులో భాగంగా ఆరోగ్యం, పర్యాటకం, వ్యవసాయం, వినోదం వంటి రంగాలలో రెండు విడతలలో 81వేల తెలుగు పాఠ్యభాషానిధి అందుబాటులో ఉంది.

టెక్నాలజీ డెవలప్మెంట్ ఫర్ ఇండియన్ లాంగ్వేజెస్ (TDIL)

ఎలక్ట్రానిక్స్ అండ్ ఇన్ఫర్మేషన్ టెక్నాలజీ మంత్రిత్వశాఖ టిడిఐఎల్ (టెక్నాలజీ డెవలప్మెంట్ ఫర్ ఇండియన్ లాంగ్వేజెస్) కార్యక్రమాన్ని ప్రారంభించింది. భాషా అవరోధాలులేని మానవ-యంత్ర పరస్పర సంభాషణలను సులభతరం చేయడానికి సమాచార విశ్లేషక సాధనాలు (ఇన్ఫర్మేషన్ ప్రాసెసింగ్ టూల్స్), సాంకేతికతను అభివృద్ధి చేయడం, బహుభాషా జ్ఞాన వనరులను సృష్టించడం, సమాచారాన్ని పొందడం, వివిధ వినియోగదారు ఉత్పత్తులు మరియు సేవలను అభివృద్ధి చేయడానికి వాటిని ఏకీకృతం చేయడం, అధికారికంగా గుర్తించబడిన 22 భారతీయ భాషలకు సాఫ్ట్ వేర్ సాధనాలను మరియు అనువర్తనాలను అభివృద్ధి చేయడం, ప్రోత్సహించడం, సృజనాత్మక ఉత్పత్తులకు, సేవలకు దారితీసే భవిష్యత్తు సాంకేతిక పరిజ్ఞానాల సహకార అభివృద్ధికి దోహదం చేయడం, భాషా సాంకేతిక ఉత్పత్తులను విస్తరించడానికి ఉత్ప్రేరకంగా పనిచేయడం, అన్ని స్థాయిలలో పరిష్కారాలు, ప్రామాణికతను అందించడం వంటివి దీని ప్రాథమిక లక్ష్యాలు. తెలుగులో 5.4 GB వాగ్భాషానిధిని సిద్ధం చేసింది.

ఇండోవర్డ్ నెట్:

భారతీయభాషలకు వర్డ్ నెట్ లను నిర్మించే అతిపెద్దప్రాజెక్టును ఇండో వర్డ్ నెట్. ఇది భారతదేశంలోని 18 షెడ్యూల్డ్ భాషలైన అస్సామీ, బంగ్లా, బోడో, గుజరాతీ, హిందీ, కన్నడ, కాశ్మీరీ, కొంకణి, మలయాళం, మీతై (మణిపురి), మరాఠీ, నేపాలీ, ఒడియా, పంజాబీ, సంస్కృతం, తమిళం, తెలుగు, ఉర్దూ భాషల 'అనుసంధానిత నైఘంటుక విజ్ఞతాధారం' (Linked Lexical Knowledge base).

కామన్ వాయిస్ (Common Voice)

కామన్ వాయిస్ అనేది వాయిస్ రికగ్నిషన్ ను అందరికీ అందుబాటులో ఉంచడంలో సహాయపడే ప్రాజెక్ట్. ఇది ప్రపంచవ్యాప్తంగా స్వచ్ఛందంగా సంభాషణలను అందించేవారి వలన నడుస్తుంది. వాయిస్ రికగ్నిషన్ టెక్నాలజీలను నిర్మించడానికి డెవలపర్లకు అపారమైన వాయిస్ డేటా అవసరం. డేటా ఎక్కువ భాగం ఖరీదైనది. కాబట్టి ఇది వాయిస్ డేటాను ఉచితంగా, బహిరంగంగా అందుబాటులో ఉంచాలనుకుంటున్నది. వాయిస్ రికగ్నిషన్ ను మరింత మెరుగ్గా చేయడం దీని లక్ష్యం. ఇక్కడ పద్దెనిమిది గంటల తెలుగు వాగ్భాషానిధి అందుబాటులో ఉంది.

ఓపెన్ ఎస్ ఎల్ ఆర్ (Open SLR)

ఓపెన్ఎస్ఎల్ఆర్ అనేది స్పీచ్, భాషా వనరులను నడపడానికి ఏర్పాటయినది. ఇది స్పీచ్ రికగ్నిషన్ కోసం ట్రైనింగ్ కార్పొరా వంటి స్పీచ్ రికగ్నిషన్ కు సంబంధించిన సాఫ్ట్వేర్. ఇది ఎక్కువ నాణ్యత కలిగిన తెలుగు మల్టీస్పీకర్ స్పీచ్ డేటాసెట్లను అందిస్తుంది. భాషా పరిశోధనను సులభతరం చేస్తుంది. ఫెయిల్ ఓవర్ స్థానాన్ని

అందించడానికి, మిగతా చోట్ల అందుబాటులో ఉన్న సాఫ్ట్ వేర్ ను తీసుకురావడం కూడా దీని లక్ష్యం. ఇది తెలుగు వాగ్భాషానిధిని కలిగి ఉంది.

ఓపెన్-స్పీచ్-ఎక్ స్టెప్/యూఎల్సీఏ- ఏఎస్ఆర్-డేటాసెట్-కార్పస్ (Open-Speech-EKSTEP/ ULCA- ASR-DATASET-COPUS)

భారతీయ భాషలలో స్పీచ్ రికగ్నిషన్, నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్ పై ఆసక్తి ఉన్నవారికి ఓపెన్ స్పీచ్ ఎక్ స్టెప్ కార్పస్ ఒక విలువైన వనరు. ఇది ఓపెన్ సోర్స్ డేటాసెట్. ఇది బహుళ భారతీయ భాషలలో వివిధ రకాల ప్రసంగ నమూనాలను కలిగి ఉంటుంది. స్పీచ్ రికగ్నిషన్ వ్యవస్థలకు శిక్షణ, మూల్యాంకనం కోసం దీనిని ఉపయోగించవచ్చు. స్పీచ్ రికగ్నిషన్ టెక్నాలజీని భారతదేశ విభిన్న భాషా ప్రదేశాలకు మరింత దగ్గరగా, అనుగుణంగా మార్చడం ఈ ప్రాజెక్టు లక్ష్యం. ఇక్కడ 1025.93 గంటల తెలుగు వాగ్భాషానిధి అందుబాటులో ఉంది.

యల్ఇఆర్సి-హెచ్.సి.యు (LERC-UOH)

హైదరాబాద్ విశ్వవిద్యాలయంలోని స్కూల్ ఆఫ్ కంప్యూటర్ అండ్ ఇన్ఫర్మేషన్ సైన్సెస్ లోని లాంగ్వేజ్ ఇంజనీరింగ్ రీసెర్చ్ సెంటర్ లో లాంగ్వేజ్ అండ్ స్పీచ్ టెక్నాలజీస్ పైన రీసెర్చ్ అండ్ డెవలప్మెంట్ గత 25 సంవత్సరాలుగా జరుగుతోంది. భారతీయ భాషలపై, ముఖ్యంగా తెలుగు (భారతదేశంలో ఎక్కువమంది మాట్లాడే రెండవ అతిపెద్ద భాష తెలుగు, ఇక్కడ స్థానిక భాష), కన్నడ (ద్రావిడ కుటుంబానికి చెందిన మరొక ప్రధాన భాష), అలాగే ఇంగ్లిష్ పై కూడా దృష్టిని కేంద్రీకరించారు. తెలుగుకు స్పీకర్ ఇండిపెండెంట్ కంటిన్యూషన్ స్పీచ్ రికగ్నిషన్ సిస్టమ్, సంబంధితరంగాలలో సమాచార వనరులు, సాధనాలు, సాంకేతికత ఉన్నాయి. ప్రస్తుతం తెలుగుకు దాదాపు 40 మిలియన్ల పదాల పాఠ్యభాషానిధి అందుబాటులో ఉన్నది.

లిప్జిగ్ కార్పొరా కలెక్షన్ (Leipzig Corpora Collection)

జర్మనీకి చెందిన లిప్జిగ్ విశ్వవిద్యాలయంలోని సాక్సన్ అకాడమీ ఆఫ్ సైన్సెస్ అండ్ హ్యుమానిటీస్, ఇన్ స్టిట్యూట్ ఫర్ అప్లైడ్ ఇన్ఫర్మేటిక్స్ సంయుక్తంగా చేపట్టిన ప్రాజెక్టు లిప్జిగ్ కార్పొరా కలెక్షన్. ఇది ఒకే ఫార్మాట్ ఉన్న, పోలికలు కలిగిన వనరులను ఉపయోగించి వివిధ భాషలలో కార్పొరాను అందిస్తుంది. మొత్తం డేటా సాదా టెక్స్ట్ ఫైళ్లుగా లభ్యం అవుతుంది. ఇక్కడ 250కి పైగా భాషలలో 900కి పైగా కార్పొరాలకు యాక్సెస్ ను అందిస్తుంది. దాదాపు 10,000 నుండి ఒక మిలియన్ వాక్యాల కార్పొరా అందుబాటులో ఉంది. ఈ సామగ్రి అంతా అంతర్జాలంలో అందుబాటులో ఉన్న వార్తాపత్రికలు, వెబ్సైట్లు, వివిధ గ్రంథాలనుంచి స్వయంచాలకంగా సేకరించబడుతుంది. తెలుగుకు 7,50,632 వాక్యాల పాఠ్యభాషానిధిని అందుబాటులో ఉంచింది.

ఐఐఐటిహైదరాబాద్ (IIIT HYD)

హైదరాబాద్ ఐఐఐటీలోని కేసీఐఎస్ లోని లాంగ్వేజ్ టెక్నాలజీస్ రీసెర్చ్ సెంటర్ లో జి.రామ రోహిత్ రెడ్డి 'సెంటిరామా' అనే కార్పస్ ను రూపొందించారు. కార్పస్ 2-వాల్యూ స్కేల్ ఉపయోగించి నాలుగు డేటాసెట్లను కలిగి ఉంటుంది. ఇది డాక్యుమెంట్ స్థాయిలో అనుకూల, ప్రతికూల సెంటిమెంట్ మధ్య వ్యత్యాసాన్ని చూపుతుంది. కార్పస్ పుస్తక సమీక్షలు, ఉత్పత్తి సమీక్షలు, సినిమా సమీక్షలు, పాటల సాహిత్యం వంటి బహుళ డొమైన్ల నుండి డేటాసెట్లను కలిగి ఉంటుంది. వాటిలో ప్రతిదాన్ని విశ్లేషకులు జాగ్రత్తగా వివరించారు. తెలుగుకు 2,98,630 పదాల పాఠ్యభాషానిధిని కలిగివుంది.

భాషాశాస్త్రశాఖహైదరాబాద్ విశ్వవిద్యాలయం:

గత 30 సంవత్సరాలుగా, హైదరాబాద్ విశ్వవిద్యాలయంలోని అనువర్తిత భాషాశాస్త్ర అనువాద అధ్యయన కేంద్రం(కాల్ట్స్), సంగణక భాషాశాస్త్ర పరిధిలో అనుసారక యంత్రానువాద పరియోజనలో భాగంగానూ, ఆ తర్వాత తెలుగూ తదితర భారతీయభాషల మధ్య యంత్రానువాదవ్యవస్థల నిర్మాణంలో భాగంగానూ జరిపిన పరిశోధనలలో భాగంగా కోటిపదాల నిడివిగల తెలుగు పాఠ్యనిధిని నిర్మించింది. తెలుగు-హిందీ, తెలుగు-తమిళం, ఇంగ్లీషు-తెలుగు యంత్రానువాద వ్యవస్థల నిర్మాణంకోసం 5 లక్షల పదాల నిడివిగలిగిన కథానికలూ, పిల్లల కథలూ, వయోజనవిద్యా సమాచార సామగ్రినుంచీ సేకరించి 15 లక్షల పదాల ప్రత్యేక పాఠ్యనిధిని తయారుచేశారు. దీనిలో భాగంగా హిందీ-తెలుగు, తెలుగు-తమిళం, ఇంగ్లీషు-తెలుగు నిఘంటువుల నిర్మాణం జరిగింది2 (చూ. cats.uohyd.ac.in).

ఇండిక్ కార్ప్ (INDICCORP)

ఇండిక్ కార్ప్ అనేది ఒక పెద్ద ఏకభాషా కార్పొరా. ఇది పన్నెండు ప్రధాన భారతీయ భాషలలో సుమారు 9 బిలియన్ టోకెన్లను కలిగి ఉంది. కేవలం తక్కువ సమయంలో వేలాది అంతర్జాల వనరులను, ప్రధానంగా వార్తలు, పత్రికలు, పుస్తకాలను తీసుకొని అభివృద్ధి చేయబడింది. అస్సామీ, బెంగాలీ, ఇంగ్లీష్, గుజరాతీ, హిందీ, కన్నడ, మలయాళం, మరాఠీ, ఒరియా, పంజాబీ, తమిళం, తెలుగు వంటి భాషలు ఇందులో ఉన్నాయి. తెలుగులో 47.9 మిలియన్ వాక్యాల పాఠ్యభాషానిధిని కలిగి ఉంది.

ఆస్కార్ (OSCAR)

ఆస్కార్ ప్రాజెక్ట్ (ఓపెన్ సూపర్-లార్జ్ క్రాల్డ్ అగ్రిగేటెడ్ కార్పస్) అనేది మెషిన్ లెర్నింగ్, ఆర్టిఫిషియల్ ఇంటెలిజెన్స్ అనువర్తనాలకోసం జాలవేదిక ఆధారంగా బహుభాషా వనరులు, డేటాసెట్లను అందించడానికి ఉద్దేశించిన ఓపెన్ సోర్స్ ప్రాజెక్ట్. మొత్తం 151 భాషలలో కార్పొరా అందుబాటులో ఉంది. తెలుగులో 13,77,52,065 పదాల పాఠ్యభాషానిధిని కలిగి ఉంది.

ఎన్పిఎల్టి (NPLT)

భారతీయ భాషా సమాచారం, సాధనాలు, సంబంధిత జాల సేవలను విద్యావేత్తలకు, పరిశోధకులకు, పరిశ్రమలకు అందుబాటులో ఉంచడానికి ఏర్పాటైన జాలవేదిక నేషనల్ ప్లాట్ఫామ్ ఫర్ లాంగ్వేజ్ టెక్నాలజీ (ఎన్పిఎల్టి). ప్రభుత్వం, స్టార్టప్ లు, పరిశ్రమలు, పోటీదారులచేత అభివృద్ధి చేయబడిన ఈ వేదిక భాషావనరులను, భాషాసాధనాలను, భాషాసేవలను అందించే మార్కెట్ లా పనిచేస్తుంది. నేచురల్ లాంగ్వేజ్ ప్రాసెసింగ్ పై ఆసక్తి ఉన్న స్టార్టప్ లు, ఎంఎస్ ఎంఈలు, ఇంటర్నేషనల్ అకడమిక్ రీసెర్చర్స్, ఎంఎన్ సీలు, విదేశీ సంస్థలు ఈ పోర్టల్ నుంచి భారతీయ భాషల వనరులను పొందవచ్చు. ఇది భారతీయ భాషాడేటా, సాంకేతికసేవలు వంటి వాటి ఆవిష్కరణలే లక్ష్యంగా పనిచేస్తున్నది. ఇక్కడ 91:49:05 గంటల తెలుగు వాగ్భాషానిధి అందుబాటులో ఉంది.

౬. సాంకేతిక వనరులు:

తెలుగుకు సాంకేతిక వనరులు ఎక్కువగానే అందుబాటులో ఉన్నాయి. లక్షలాది పుస్తకాలు, వ్యాసాలు, మాసపత్రికలు, నిఘంటువులు వివిధ జాలవేదికల్లో ఉచితంగా అందుబాటులో ఉన్నాయి. కేవలం భారతీయ సంస్థలేకాక విదేశీసంస్థలు కూడా ఈ దిశగా కృషి చేయటాన్ని గమనించవచ్చు. మార్కెట్లో భాషలకోసం అందుబాటులోకి వస్తున్న నూతన ఆవిష్కరణలన్నిటిలో తెలుగు భాష కూడ ఉండటం గమనార్హం. కృత్రిమమేధకు సంబంధించిన టూల్స్, అప్లికేషన్స్ అన్నిటిలో తెలుగు తప్పనిసరిగా ఉంటుంది.

భారతవాణి (BHARATVANI)

సెంట్రల్ ఇనిస్టిట్యూట్ ఆఫ్ ఇండియన్ లాంగ్వెజెస్, మైసూరు ఈ ప్రాజెక్టును అమలు చేస్తోంది. ఆన్లైన్ పోర్టల్ ద్వారా మల్టీమీడియా (టెక్స్ట్, ఆడియో, వీడియో, ఇమేజెస్) ఫార్మాట్లలో భారతదేశంలోని అన్ని భాషల్లో శోధించదగిన జ్ఞాన భాండాగారాన్ని నిర్మించాలన్నది భారతవాణి లక్ష్యం. భారతవాణి పోర్టల్ సమాజంలోని అన్ని వర్గాలకు అందుబాటులో ఉంటుంది. కాపీరైట్ (సవరణ) చట్టం, 2012 యొక్క న్యాయమైన వినియోగ క్లాజుల కింద ఈ పోర్టల్ విద్యా ప్రయోజనాల కోసం ఓపెన్ నాలెడ్జ్ ను అందిస్తుంది. ఇక్కడ తెలుగులో అనేక నిఘంటువులు, భాషాగ్రంథాలు, సాంస్కృతిక గ్రంథాలు అందుబాటులో ఉన్నాయి.

ఆంధ్రభారతి (ANDHRABHARATI)

అంతర్జాలంలో సెర్చ్ ఇంజిన్ ఉన్న ప్రముఖ తెలుగు నిఘంటువులలో AndhraBharati.com ఒకటి. ఈ వెబ్ సైట్ ను వాడపల్లి శేషాచలసాయి, కాలెపు నాగభూషణరావు నిర్వహిస్తున్నారు. ఇక్కడ తెలుగు సాహిత్యంలోని వివిధ సాహితీ ప్రక్రియలకు సంబంధించిన పాఠం టైప్ చేయించి అందుబాటులో ఉంచారు. ఇది నకలు చేసుకోవటానికి అనుకూలమైనది. పరిశోధకులకు ఎంతగానో ఉపకరిస్తుంది. ఇక్కడి ప్రత్యేక విభాగం 'నిఘంటుశోధన'. ఇందులో 16 తెలుగు భాషా నిఘంటువులు ఉన్నాయి. మొత్తం నిఘంటువుల సంఖ్యను 71కి

పెంచే ప్రయత్నానికి 'తానా' సంస్థ సహకరిస్తోంది. ఇక్కడ కేవలం తెలుగు నిఘంటువులేకాక సంస్కృత నిఘంటువులు కూడా అందుబాటులో ఉన్నాయి. ఒక పదాన్ని నిర్దేశిత స్థలంలో టైపుచేసి మీట నొక్కితే ఆ పదం ఏఏ నిఘంటువులలో ఉందో ఆ నిఘంటువులన్నీ అర్థాలను అందిస్తాయి. ఇలా ఆంగ్లంలో ఇచ్చిన పదానికి తెలుగులో, తెలుగులో ఇచ్చిన పదానికి ఆంగ్లంలో ఆర్థాలను అందిస్తుంది. ఇలా ఒకే వేదిక మీద అన్ని నిఘంటువులు అందుబాటులో ఉంచడం ఈ సంస్థ ప్రత్యేకత. ఇది పాఠకునికి ఎంతో ప్రయోజనకరం. ఇలాంటి సౌకర్యాల ఆధారంగా ఉత్తమ తెలుగు సాహిత్యాన్ని సాహితీ ప్రియులకు, పరిశోధకులకు అందుబాటులో ఉంచాలనే గొప్ప లక్ష్యంతో ఈ సంస్థ పనిచేస్తున్నది.

శోధ్ గంగ:

శోధ్ గంగ భారతీయవిశ్వవిద్యాలయాలలో జరిగిన పరిశోధనసిద్ధాంతవ్యాసాల రిజర్వాయర్. భారతీయవిశ్వవిద్యాలయాలలో జరిగిన పరిశోధనలను ఎలక్ట్రానిక్ రూపంలో ఇక్కడ నిక్షిప్తం చేశారు. దీనిని INFLIBNET కేంద్రం నడిపిస్తుంది. ఇది వివిధ భారతీయభాషలలో పరిశోధనలు చేయదలచుకున్న వారికి, నూతనపరిశోధకులకు గొప్ప వనరు. పరిశోధన దిక్సూచి. విశ్వవిద్యాలయాలలో జరిగే పరిశోధనలు పునరావృతంకాకుండా ఉండటానికి, పరిశోధన ప్రమాణాలను, నాణ్యతను పెంపొందించే లక్ష్యంతో ఇది ఏర్పాటుచేయబడింది. యుజిసి ఆదేశాల ప్రకారం భారతీయ విశ్వవిద్యాలయాలలో జరిగే యం.ఫిల్, పిహెచ్.డి పరిశోధనల ఎలక్ట్రానిక్ వెర్షన్ ను ఇక్కడ పొందుపర్చాలి. ఇక్కడ మొత్తం 590375 సిద్ధాంతవ్యాసాలు, 15462 సిద్ధాంతవ్యాససారసంగ్రహాలు, 83 ఫెలోషిప్ లకు సంబంధించిన వ్యాసాలు అందుబాటులో ఉన్నాయి. దేశవ్యాప్తంగా తెలుగులో వచ్చిన సిద్ధాంతవ్యాసాలు కొంత వరకు ఇక్కడ అందుబాటులోకి వచ్చాయి.

డిజిటల్ లైబ్రరీ ఆఫ్ ఇండియా:

ఇంటర్నెట్ ఆర్కైవ్, డీజిటల్ లైబ్రరీ ఆఫ్ ఇండియా బహుశా ఈ పేరు తెలియని ఈ తరం పరిశోధకులు ఉండరంటే ఆశ్చర్యం లేదు. ఇది లాభాపేక్షలేని ఇంటర్నెట్ ఆర్కైవ్, డిజిటల్ రూపంలో ఇంటర్నెట్ సైట్లు, ఇతర సాంస్కృతిక కళాఖండాల డిజిటల్ లైబ్రరీని నిర్మిస్తోంది. పేపర్ లైబ్రరీ వలె, పరిశోధకులకు, చరిత్రకారులకు, పండితులకు, సాధారణ ప్రజలకు ఉచితంగా పుస్తకాలను పొందటానికి అవకాశం కల్పిస్తుంది. సర్వజ్ఞానాన్ని సార్వత్రికంగా అందించడమే దీని లక్ష్యం.

దీనిని 1996 లో ప్రారంభించారు. అప్పటినుంచే ఇది అందుబాటులోకి రావడం ప్రారంభమైంది. వార్తాపత్రికల వలె, వెబ్ లో ప్రచురించబడిన కంటెంట్ తాత్కాలికమైనది. కానీ వార్తాపత్రికల మాదిరిగా, ఎవరూ దానిని సేవ్ చేయడం లేదు. ఈ రోజు మనకు వేబ్యాక్ మెషిన్ ద్వారా 29 సంవత్సరాల వెబ్ చరిత్ర అందుబాటులో ఉంది. ముఖ్యమైన వెబ్ పేజీలను గుర్తించడానికి ఇది ఆర్కైవ్-ఇట్ ప్రోగ్రామ్ ద్వారా 1,200 పైచిలుకు లైబ్రరీ, ఇతర భాగస్వాములతో కలిసి పనిచేస్తున్నది.

ప్రస్తుతం 835 బిలియన్ వెబ్ పేజీలు, 44 మిలియన్ల పుస్తకాలు, 15 మిలియన్ ఆడియో రికార్డింగ్ లు (255,000 లైవ్ కచేరీలతో సహా), 10.6 మిలియన్ వీడియోలు (2.6 మిలియన్ టెలివిజన్ న్యూస్ కార్యక్రమాలతో సహా), 4.8 మిలియన్ చిత్రాలు, 1 మిలియన్ సాఫ్ట్ వేర్ ప్రోగ్రామ్ లు అందుబాటులో ఉన్నాయి.

దక్షిణాసియా డిజిటల్ నిఘంటువులు (DDSA)

డిజిటల్ సౌత్ ఆసియా లైబ్రరీ అనేది ఆండ్రూ డబ్ల్యూ మెల్లన్ ఫౌండేషన్ మద్దతుతో అసోసియేషన్ ఆఫ్ రిసెర్చ్ లైబ్రరీస్ గ్లోబల్ రిసోర్సెస్ ప్రోగ్రామ్ ద్వారా నిధులు సమకూర్చబడిన రెండు సంవత్సరాల పైలట్ ప్రాజెక్టు ఆధారంగా రూపొందించబడింది. ఇది పండితులకు, పరిశోధకులకు, అధికారులకు, వ్యాపారులకు, దక్షిణాసియాపై పరిశోధనలు చేసే ఇతర ఏ వినియోగదారులకైనా డిజిటల్ మెటీరియల్ ను అందిస్తుంది. ఇందులో భారతదేశంలోని సంస్థలతోపాటు ప్రపంచవ్యాప్తంగా ఉన్న బ్రిటీష్ లైబ్రరి, యూనివర్సిటీ ఆఫ్ కేంబ్రిడ్జ్, యూనివర్సిటీ ఆఫ్ ఆక్స్ ఫర్డ్ వంటి ప్రముఖ సంస్థల భాగస్వామ్యం ఉంది. ద్రావిడ భాషలపై పరిశోధనల అభివృద్ధి దీని లక్ష్యం.

లెక్సిలోగోస్ (LEXILOGOS)

లెక్సిలోగోస్ అనేది ప్రపంచంలోని భాషల అధ్యయనానికి ఒక సమగ్ర వనరుల సమూహం. లెక్సిలోగోస్ గ్రీకు భాషాపదం. ఇది నిఘంటువును అధ్యయనం చేసే శాస్త్రం. ఆధునిక గ్రీకు భాషలో నిఘంటువు అని అర్థం. ప్రపంచంలోని ఇతర ప్రాంతాలు, ఇతర సంస్కృతుల భాషలను అన్వేషించడం ద్వారా మన పదజాలాన్ని సుసంపన్నం చేద్దాం అన్నది దీని లక్ష్యం. ఇక్కడ వివిధ భాషలకు చెందిన నిఘంటువులు అందుబాటులో ఉన్నాయి. తెలుగులో ఉన్న ప్రసిద్ధ నిఘంటువులు ఇక్కడ అందుబాటులో ఉన్నాయి. ఆంధ్రభారతి నిఘంటుశోధన కూడా ఇక్కడ అందుబాటులో ఉంది. తెలుగులో వచ్చిన భాషాశాస్త్ర గ్రంథాలు, తెలుగుపై వచ్చిన ఆంగ్లగ్రంథాలు, వ్యాసాలు, వివిధ తెలుగు మాండలికాలకు సంబంధించిన సంభాషణలు ఆడియో రూపంలో అందుబాటులో ఉన్నాయి.

గ్లోస్బే (GLOSBE)

గ్లోస్బే అతిపెద్ద కమ్యూనిటీ-నిర్మిత నిఘంటువు. ఇది ప్రపంచంలోని అన్ని భాషలను సపోర్ట్ చేస్తుంది. గ్లోస్బే అనేది సందర్భోచిత అనువాదాలతో (అనువాద వాక్యాలు - అనువాద జ్ఞాపకశక్తి అని పిలువబడేది) ఉచిత నిఘంటువులను అందించే వేదిక. ఇక్కడ తెలుగు నుంచి ప్రపంచభాషలకు మాటలు ఆడియో రూపంలో అందుబాటులో ఉన్నాయి.

కేంబ్రిడ్జ్ నిఘంటువు (Cambridge Dictionary)

కేంబ్రిడ్జ్ డిక్షనరీలో తెలుగు నుంచి ఆంగ్లం, ఆంగ్లం నుంచి తెలుగు వంటివి అందుబాటులో ఉన్నాయి. దీనితోపాటు అనువాద వేదిక, గ్రామర్ చెక్కర్స్, థెసారస్ వంటివి అందుబాటులో ఉన్నాయి.

శబ్దకోశ్.కమ్ (SHABDKOSH.COM)

ఆంగ్లం నుంచి భారతీయ భాషలకు నిఘంటు సేవలను అందిస్తున్నది. ఉపయోగించడానికి సులభమైన ఇంటర్ఫేస్, సమగ్ర డేటాబేస్, బహుళ ఉచ్చారణలలో వాయిస్ ఉచ్చారణలు వంటి ఉపయోగకరమైన లక్షణాలతో, భారతీయ భాషా వనరులు ప్రపంచంలోని మరే ఇతర భాషకు లేనంత మెరుగ్గా ఉన్నాయని నిర్ధారించడమే లక్ష్యంగా పనిచేస్తుంది.

ఇండియాడిక్ట్ (INDIADICT)

ఇది కొన్ని సెకన్ల కంటే తక్కువ సమయంలో ఒక ఆంగ్ల పదానికి తెలుగు అర్థాన్ని అందిస్తుంది. భారతదేశంతోపాటు ప్రపంచవ్యాప్తంగా లక్షలాది మంది ఇంగ్లీష్ మాట్లాడే ప్రజలు ఇంగ్లీష్ నుండి తెలుగు ఆన్లైన్ నిఘంటువు కోసం చూస్తున్నారు. కాబట్టి, ఇండియాడిక్ట్ మనకు ఉచితంగా ఆంగ్లం నుండి తెలుగు నిఘంటువును అందిస్తుంది. మనం ఇంగ్లీష్ పదాన్ని టైప్ చేస్తే ఇండియాడిక్ట్ మనకు ఆ పదానికి తెలుగు అర్థాన్నిస్తుంది. ఇక్కడ ఇంగ్లీషు-తెలుగు నిఘంటువులో లక్షలాది పదాలు, వాటి అర్థాలు ఉన్నాయి.

పై జాలవేదికల్లో వందలాది నిఘంటువులు, లక్షలాది పీడియఫ్ పుస్తకాలు అందుబాటులో ఉన్నాయి. ఇవికాక టిటిడి తిరుపతి, సుందరయ్య విజ్ఞానకేంద్రం, తెలుగువన్ గ్రంథాలయం, కౌముది గ్రంథాలయం, ఈ పుస్తకాలు - తెలుగు పుస్తకాలు, తెలుగుపీడిఎఫ్, ఓపెన్ లైబ్రరీ, తెలుగుభారత్, తెలుగు పరిశోధన, తెలుగుతల్లి సేవలో.., ఉచిత గురుకుల విద్యా ఫౌండేషన్, కథానిలయం, శ్రీ రామకృష్ణ సేవాసమితి, సాధకుడు వంటి అనేక రకాల తెలుగు పీడియఫ్ పుస్తకాల వేదికలు ఉచితంగా అందుబాటులో ఉన్నాయి. ప్రపంచంలో ఏ మూలనున్న తెలుగు పరిశోధకుడికైనా లేదా తెలుగు పై పరిశోధిస్తున్న ఏ ఇతర భాషీయుడికైనా ఈ నిఘంటువులు, పుస్తకాలు అందుబాటులో ఉన్న గొప్ప వనరులు. భాషల్ని పరిరక్షిస్తున్న ఈ సంస్థలన్నీ పదజాలాన్ని దినదినాభివృద్ధి చేయటంతోపాటు భాషను విశ్వవ్యాప్తం చేస్తున్నాయి.

ఈ క్రమంలోనే తెలుగులో అందుబాటులో ఉన్న యంత్రానువాద వేదికలు, జెనరిక్ ఆర్టిఫిషియల్ ఇంటిలిజెన్స్, ఓసియార్ టూల్స్, టైపింగ్ కీబోర్డ్స్ వంటి వాటిని గురించి చెప్పుకోవాలి. ప్రస్తుతం ఎఐ టూల్స్ పదుల సంఖ్యలో అందుబాటులో ఉన్నాయి. నిజానికి ఏఐ టూల్స్ కు సంబంధించి ఉదాహరణలతో కూడిన వివరణాత్మక వ్యాసం అవసరం. ఇక్కడ సంక్షిప్త సమాచారం మాత్రమే అందిస్తున్నాను.

యంత్రానువాదంలో మెషిన్ ట్రాన్సులేషన్.కమ్ ఒక ప్రత్యేకమైన వేదిక. ఇది చాట్జిపిటి, మైక్రోసాఫ్ట్, క్లౌడ్, జెమిని, అమెజాన్, లింగ్వెనెక్స్, ఫేస్ బుక్ వంటి వేదికల సమ్మేళనం. ఈ అనువాద వేదికలో ఒక వాక్యాన్ని ఒక భాషనుంచి ఇంకో భాషకు అనువదించడానికి సూచించినట్లైతే అది చాట్జిపిటి, మైక్రోసాఫ్ట్, క్లౌడ్, జెమిని, అమెజాన్, లింగ్వెనెక్స్, ఫేస్ బుక్ వంటి టూల్స్ ఈ వేదికపై ఒకేసారి వాటి అనువాదాలను అందిస్తుంది. ఇక్కడ మనకు నచ్చిన అనువాదాన్ని ఎంచుకొని ఆ అనువాదాన్ని కాపీ చేసుకోవచ్చు, డౌన్లోడ్ చేసుకోవచ్చు లేదా ఈ అనువాద లింకును వేరెవరికైనా పంపవచ్చు. డౌన్లోడ్ బటన్ పై నొక్కితే ఆ అనువాదం ప్రత్యేకమైన ఒక వర్డ్ ఫైల్ లోకి దిగుమతి అవుతుంది. ఈ టూల్ ద్వారా ఒకేసారి మూడువేల ఆంగ్ల పదాలను అనువాదం చేసుకోటానికి

వీలుంది. మిగతా అనువాదాలను కూడా ఒకే వేదికపై అందించే ఈ టూల్ ఎంతో ఉపయోగకరంగా ఉంటుంది. ఇలాంటి అనువాద వేదికల్లో గూగుల్ ట్రాన్స్లేషన్, మైక్రోసాఫ్ట్ బింగ్, భాషిణి, ఎల్డీసిఐయల్ అనువాదిక, వర్డ్ వాయిస్.ఎఐ, హెచ్ ఐయక్స్.ఏఐ, తెలుగు ఇండియా టైపింగ్, ట్రాన్స్లేట్ పల్స్, శబ్దకోశ్, టైపింగ్ బాబా, ట్రాన్స్లేషన్లి, అనువాదిని, కేంబ్రిడ్జ్ డిక్షనరీ, ఆన్లైన్ డాక్ ట్రాన్స్లేటర్ వంటి అనేక వేదికలున్నాయి.

ఛాట్ జిపిటి, కోపైలెట్, ఓపెన్ ఏఐ, డీప్ సిక్, ఓలా కృత్రిమ్, భారత్ జెన్, మేటా, జెమిని వంటివి అడ్వాన్సుడ్ ఏఐ లాంగ్వేజ్ మోడల్స్ ప్రస్తుతం మార్కెట్లో ఉన్నాయి. ఇవి తెలుగులో మనం అడిగిన ప్రశ్నలకు తెలుగులోను, ఒకవేళ ఇతర భాషలలో సమాధానం కోరినా చెప్పటానికి ప్రయత్నిస్తాయి. ఏ రంగానికి సంబంధించిన అందుబాటులో ఉన్న సమాచారాన్నైనా అందిస్తాయి. మేటా వంటి వాటిలో తెలుగు వంటి భారతీయభాషలు కొన్ని ఇంకా సపోర్టు చేయవలసి ఉంది. ఈ జనరిక్ ఏఐ రోజురోజుకు తన పరిధిని విస్తృతం చేసుకుంటుంది. కొన్ని సందర్భాల్లో తననుతాను సరిచేసుకోవటానికి కావలసిన సమాచారాన్ని సంబంధిత రంగం నుంచి అందించగలిగితే మరోసారి ఖచ్చితమైన సమాధానాన్ని అందించటానికి ప్రయత్నిస్తుంది.

7. ఉపసంహారం:

ఇది సాంకేతిక ప్రపంచం, ఏఐ ప్రపంచం. కృత్రిమమేధ మానవమేధస్సుకు సమానంగా పోటీ పెడుతున్న కాలమిది. యంత్రాలకు కృత్రిమమేధను అందించే సందర్భంలో భాష ప్రధానమైనది. భాషలో సంభాషణలు మరింత ముఖ్యమైనవి. ఈ సంభాషణలకు భాషలోని టెక్స్ట్, స్పీచ్ డేటాసెట్లు అత్యంత అవసరం. ఈ డేటాసెట్ల తయారీ అంతసులభమైన పనికాదు. డబ్బుతోపాటు సమయం, మానవమేధస్సు, వివిధ వయసుల వ్యక్తుల సంభాషణలు, ముద్రిత పాఠ్యం వంటివెన్నో అవసరం. వాటి విశ్లేషణలు కూడా ఒక ప్రత్యేక అంశం. అలాగే ముద్రిత గ్రంథాలను డిజిటలైజ్ చేయటం కూడా డబ్బు, కాలం, శ్రమతో కూడిన అంశమే. ఇలా తెలుగు భాషాభివృద్ధి కోసం స్వచ్ఛందంగా ప్రభుత్వ, ప్రైవేటు సంస్థలు శ్రమిస్తున్నాయి. ఇది వివిధ సంస్థల ప్రణాళికలతో కూడిన సమిష్టి ప్రయత్నం. ఈ ప్రయత్నాలు మరింత మెరుగైనవిగా, ఖచ్చితత్వంతో కూడినవై భాషాభివృద్ధికి మరింత ఉపకరించాలని అందరూ కోరుకోవాలి. అయితే ఈ సంస్థలు రూపొందించిన పరికరాల పనితీరును పరిశీలించటం కూడా ఒక ప్రత్యేక అధ్యయనమే. మొత్తానికి ఈ వనరులు భాషా సాంకేతిక పరిజ్ఞానం పురోగతికి తోడ్పడటమే కాకుండా, తెలుగు భాషా వారసత్వాన్ని పరిరక్షించి, ప్రోత్సహిస్తాయని కచ్చితంగా చెప్పవచ్చు.

8. పాదసూచికలు:

1. https://data.ldcil.org/
2. అమ్మనుడి మాసపత్రిక, ఫిబ్రవరి-2023, పుట.29

౯. ఉపయుక్తగ్రంథసూచి:

1. రాధాకృష్ణ బూదరాజు. (2008), ఆధునిక వ్యవహారకోశం ఇంగ్లీషు - తెలుగు, ప్రాచీపబ్లికేషన్స్, హైదరాబాద్
2. రామారావు చేకూరి. (2004), పత్రికాపదకోశం ఇంగ్లీషు-తెలుగు, ఆంధ్రప్రదేశ్ ప్రెస్ అకాడమీ, హైదరాబాద్
3. కాశీంబాబు మోదుగు, రాజేశ యన్., మానస సి., నారాయణ్ ఛౌదరి, శైలేంద్రమోహన్., (2025), ప్రామాణిక తెలుగు పాఠ్యదత్తాంశం (సం.2), భారతీయ భాషాసంస్థ, మైసూర్
4. డా. రమేశ్ బాబు సామల (సంపా.) (2023), అమ్మనుడి మాసపత్రిక (ఫిబ్రవరి), తెలుగుజాతి (ట్రస్టు) ప్రచురణ, విజయవాడ
5. రామమూర్తి యల్., నారాయణ ఛౌదరి, తిరుపాల్ సి రెడ్డి, గంగరాజు హెచ్, (2019), ప్రామాణిక తెలుగు పాఠ్యదత్తాంశం, భారతీయ భాషాసంస్థ, మైసూర్
6. రామమూర్తి యల్., నారాయణ ఛౌదరి., రాజేశ యన్., (2019), తెలుగు సహజ వాగ్దత్తాంశం, భారతీయ భాషాసంస్థ, మైసూర్
7. Niladri Sekhar Dash, Pushpak Bhattacharyya, Jyoti D. Pawar (Ed.), (2017), The WordNet in Indian Languages, Springer, Singapore
8. Language Technical Resources for Telugu (2025), Linguistic Data Consortium for Indian Languages, Central Institute of Indian Languages, Mysore

***

గమనిక: ఈ పత్రికలోని వ్యాసాలలో అభిప్రాయాలు రచయితల వ్యక్తిగతమైనవి.
వాటికి సంపాదకులు గానీ, పబ్లిషర్స్ గానీ ఎలాంటి బాధ్యత వహించరు.