INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

ILDDAC-2026

The Art and Science of Data Annotation for AI

Organised by

NEXTGENVECTORA DATA INNOVATIONS (OPC) PRIVATE LIMITED

Proudly supported by our valued sponsors and partners

INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

ILDDAC-2026

Organised by

NEXTGENVECTORA DATA INNOVATIONS (OPC) PRIVATE LIMITED

Proudly supported by our valued sponsors and partners

KLS Gogte Institute of Technology, Belagavi
*Permanently Affiliated and an Autonomous Institute Under Visvesvaraya Technological University, Belagavi | Accredited by NAAC A+ and NBA**
Estd. 1939
Estd. 1979

DAYANANDA SAGAR COLLEGE OF ENGINEERING
AN AUTONOMUS INSTITUTE
AFFLIATED TO VTU
APPROVED BY AICTE & UGC
ACCREDITED BY NAAC WITH 'A' GRADE
Department of Computer Science & Engineering (Data Science)

C. BYRE GOWDA INSTITUTE OF TECHNOLOGY, KOLAR
C.B.I.T.

NIGHAN2 LABS .....

SILVER MEMORIES GOLDEN MOMENTS
25
RLJIT
SILVER JUBILEE
2001-2025

SJCIT

PARITRANAYA IAS ACADEMY

B. N. M. Institute of Technology
An Autonomous Institution under VTU. Approved by AICTE.
Vidyayāmruthamashnuthe

EAST WEST INSTITUTE OF TECHNOLOGY
1968

INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

What are you actually doing?

You are teaching AI your native language.

What skill are you building?

Teaching machines to understand humans.

Is this only a competition task?

No — it’s a contribution to future AI.

How should you annotate?

With care, consistency, and pride.

One Language, Many Dialects

A language is not uniform — it lives through its dialects. Dialects reflect geography, culture, identity, and social interaction.

Kannada

- Mysore Kannada
- North Karnataka Kannada
- Coastal Kannada (Mangalore)
- Havyaka Kannada
- Bengaluru Kannada

Telugu

- Coastal Andhra Telugu
- Telangana Telugu
- Rayalaseema Telugu
- Hyderabad Telugu

Tamil

- Central Tamil – Spoken mainly in Tamil Nadu central districts (Thanjavur, Trichy belt)
- Madurai Tamil – Southern Tamil Nadu region
- Kongu Tamil – Western Tamil Nadu (Coimbatore, Erode, Salem)
- Tirunelveli Tamil – Deep south districts
- Chennai Tamil (Madras Bashai)
- Northern Tamil – Vellore–Chengalpattu region

Marathi

- Varhadi
- Khandeshi
- Malvani
- Kolhapuri Marathi

Bengali

- Rarhi Bengali – Standard Kolkata-based dialect
- Manbhumi Bengali – Western border areas
- Varendri Bengali – North Bengal region

INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

Domains to Teach AI your Native Language

(Vocabulary • Sentence • Paragraph Annotation Sources)

🏠 1. Daily Life & Home Conversations

What it teaches AI

- Natural vocabulary
- Informal sentence structure
- Dialect expressions

Examples

- Family conversations
- Household instructions
- Elders speaking to children

➡️ **This is where mother-tongue truly exists.**

🏘️ 2. Local Community & Social Interactions

What it teaches AI

- Regional dialect markers
- Politeness, respect, authority

Examples

- Talking to shopkeepers
- Neighbour conversations
- Village / town interactions

🎓 3. Education & Learning

What it teaches AI

- Structured language
- Formal vs informal contrast

Examples

- Textbooks
- Classroom instructions
- Teacher–student conversations

🏥 4. Health & Public Services

💼 5. Work & Occupation

🎭 6. Culture and Stories

📱 7. Digital & Social Media Language

⚖️ 8. Governance, Law & Public Communication

9. Agriculture & Farming

10. Health & Public Services

11. Food & Cooking

12. Sports & Recreation

Note: The domains listed above are indicative and not exhaustive. Participants are free to choose any relevant domain that reflects natural usage of the native language or dialect.

NLP Capabilities Enabled by Your Annotations

Your annotations help AI perform:

- **Question Answering –** finding correct answers from text
- **Text Summarisation –** capturing key ideas accurately
- **Topic Identification –** understanding what text is about
- **Information Extraction –** identifying names, places, actions
- **Intent Detection –** understanding user purpose
- **Sentiment Understanding –** detecting emotion and opinion
- **Dialect & Variation Handling –** understanding real native speech

Important Note

- **Poor annotations break these capabilities**
- **Careful annotations make them possible**

INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

🏭 Industry Use-Cases for Dialect Annotations

🗣️ 1. Regional Language Chatbots

- Government service portals
- Utility and grievance systems
- University and admission helpdesks

➡️ Dialect-aware bots understand how people actually speak, not textbook Kannada.

📞 2. Voice Assistants & IVR Systems

- Telecom customer care
- Banking helplines
- Rural and semi-urban support systems

➡️ Dialect annotation enables accurate speech understanding and intent detection.

🏥 3. Healthcare Communication Systems

- Patient symptom collection
- Appointment booking
- Medical instruction delivery

➡️ Dialects improve understanding of non-formal patient language.

🏛️ 4. Government & Public Welfare Platforms

- Scheme eligibility queries
- Farmer support systems
- Citizen feedback analysis

➡️ Inclusive AI for regional and rural populations.

🛒 5. Local Commerce & Customer Support

- E-commerce regional support
- Small business chat support
- Order and complaint handling

➡️ Dialect understanding improves customer trust and resolution speed.

🏦 9. Banking & Financial Inclusion

- Loan assistance
- Insurance support
- Financial literacy platforms

➡️ Dialect-aware NLP reduces miscommunication and drop-offs.

Note: These use-cases are examples only. The applications of dialect annotations extend to many other industry scenarios.

Annotation Examples in Practice

Organised by

NEXTGENVECTORA DATA INNOVATIONS (OPC) PRIVATE LIMITED

Proudly supported by our valued sponsors and partners

VOCABULARY-LEVEL ANNOTATION (TOKEN-LEVEL)

Objective:

Create dialect vocabulary dataset for:

- Dialect adaptation models
- Chatbots
- Semantic search
- Embedding training

EXAMPLES:

Textbook Kannada	North Karnataka Dialect
ಬರುತ್ತೇನೆ	ಬರ್ತೀನಿ
ಏಕೆ	ಯಾಕ

✅ Annotation Tip:

Meaning is preserved, spelling reflects **spoken North Karnataka usage**.

NLP/AI Perspective:

👉 Vocabulary should supports:

- Tokenizer building
- Embedding training
- Named entity detection
- Dialect normalization

Show your Creativity:

By including:

- Rare dialect words
- Cultural expressions
- Idioms
- Age-group variations

Tips for Participants

Include Variation Types:

A. Synonyms

- Include all the dialect synonym mapping separately

B. Slang & Informal Words

- Very valuable for LLMs.
  Example:
  ಊಟ → ತಿಂಡಿ

C. Domain Vocabulary

Encourage:

- Agriculture
- Education
- Festivals
- Technology
- Daily life

VOCABULARY-LEVEL ANNOTATION (TOKEN-LEVEL)

EXAMPLES:

Textbook Tamil	Madurai/Chennai spoken
எப்படி	எப்டி

Textbook Kannada	North Karnataka Dialect
ಬರುತ್ತೇನೆ	ಬತೀನಿ

Textbook Telugu	Telangana Telugu
ఎలా	ఎలాగ

Textbook Marathi	Vidarbha Marathi
खूप	लई

SENTENCE-LEVEL ANNOTATION

Why Only Vocabulary-level Annotation is not sufficient to train Large Language Models?

- Loses context, grammar, and natural usage patterns
- Cannot capture idioms, tone, or sentence intent
- Limited usefulness for real NLP applications
- Models fail in translation, summarisation, and conversation tasks

Importance of Sentence Annotation

- Captures grammar, context, and dialect variations
- Improves NLP tasks like translation, sentiment analysis, QA
- Helps train conversational and generative AI models
- Reflects real-world language usage

SENTENCE-LEVEL ANNOTATION

Tips for Sentence-level Annotation

- Use different sentence forms such as declarative, interrogative, imperative, and exclamatory to capture diverse linguistic patterns.
- Include both formal textbook language and informal dialect usage to reflect real-world communication.
- Cover simple, compound, and complex sentence structures for better language modeling.
- Add sentences expressing questions, commands, emotions, and opinions to improve conversational AI training.
- Incorporate idioms, slang, and culturally specific expressions to capture natural language richness.
- Select examples from multiple domains like education, agriculture, technology, daily life, and culture.
- Include variations in politeness levels (formal vs casual speech) common in dialects.
- Prefer context-rich sentences that help train NLP models for tasks such as translation, summarisation, sentiment analysis, and dialogue systems.

SENTENCE-LEVEL ANNOTATION

EXAMPLES:

Textbook Kannada	North Karnataka Dialect
ನಾನು ಇಂದು ಶಾಲೆಗೆ ಹೋಗುವುದಿಲ್ಲ.	ನಾ ಇವತ್ತ ಶಾಲಿಗ್ ಹೋಗಲ್ಲ.

Textbook Tamil	Madurai Spoken Tamil Dialect
நீங்கள் சாப்பிட்டீர்களா?	நீங்க சாப்பிட்டிங்களா?

Textbook Telugu	Telangana Telugu Dialect
మీరు ఎక్కడికి వెళ్తున్నారు?	మీరెక్కడికి వెల్తున్నరు?

Textbook Marathi	Varhadi Marathi Dialect
आज खूप गरम आहे	आज फार गरम हाय

PARAGRAPH-LEVEL ANNOTATION

Why Vocabulary & Sentence Annotation Alone is Not Enough!

Drawbacks of Only Vocabulary Annotation

- No context → words may change meaning in sentences/paragraphs
- Cannot capture grammar, tone, or discourse flow
- Limited usefulness for conversational AI and LLM training

Drawbacks of Only Sentence Annotation

- Sentences lack broader conversational or narrative context
- Cannot model long-context understanding
- Difficult to train summarisation, topic modeling, or document classification models
- Poor performance in real-world NLP applications

👉 **Key Takeaway:** Real language usage happens in paragraphs, conversations, and narratives — not isolated words or sentences.

PARAGRAPH-LEVEL ANNOTATION

Importance of Paragraph Annotation

Why Paragraph Annotation Matters

- Captures **context, discourse flow, and meaning continuity**
- Helps AI understand **natural dialect variations in real usage**
- Enables training for advanced NLP capabilities

👉 **Key Takeaway:** Paragraph-level annotation creates realistic training data for Generative AI systems.

NLP Tasks Supported

- Long-context retention in LLMs
- Text summarisation
- Text classification
- Sentiment analysis
- Topic modeling
- Named Entity Recognition (NER)
- Question answering systems
- Conversational AI systems
- Creative text generation (stories, poems, dialogues)

PARAGRAPH-LEVEL ANNOTATION

Tips for Paragraph Annotation

Include Variety of Paragraph Types

- Conversations (daily life, market, classroom, office)
- Narratives (stories, personal experiences)
- Informational paragraphs (education, agriculture, technology)
- Opinionated text (reviews, social issues)
- Emotional expressions (festivals, family events)

Include Linguistic Diversity

- Formal vs informal dialect usage
- Cultural references and idioms
- Domain-specific vocabulary
- Natural spoken flow

NLP-Focused Tips

Ensure paragraphs support:

- Context retention across sentences
- Summarisation tasks
- Sentiment detection
- Topic classification
- Entity extraction
- Question answering
- Creative generation training

👉 Always mention dialect region name during annotation.

PARAGRAPH-LEVEL ANNOTATION

EXAMPLES:

Textbook Kannada	North Karnataka Dialect
ನಾನು ಪ್ರತಿದಿನ ಬೆಳಗ್ಗೆ ಬೇಗ ಎದ್ದು ಕೆಲಸಕ್ಕೆ ಹೋಗುತ್ತೇನೆ. ಕೆಲಸ ಮುಗಿಸಿದ ನಂತರ ಮನೆಗೆ ಬರುತ್ತೇನೆ. ನನಗೆ ನನ್ನ ಕೆಲಸ ಬಹಳ ಇಷ್ಟ.	ನಾ ಪ್ರತಿದಿನ ಬೆಳಗ್ಗೆ ಬೇಗ ಎದ್ದು ಕೆಲ್ಸಕ್ಕ ಹೊಗ್ತೀನಿ. ಕೆಲ್ಸ ಮುಗ್ಸಿ ಮನೆಗ ಬರ್ತೀನಿ. ನಂಗೆ ನನ್ ಕೆಲ್ಸ ಬಾಳ ಇಷ್ಟ.

Textbook Telugu	Telangana Telugu Dialect
ఈ రోజు మార్కెట్‌లో చాలా జనసందడి ఉంది. రైతులు తమ పంటలను అమ్ముతున్నారు.	ఇవాళ మార్కెట్లో చాల రద్దీ ఉంది. రైతులు తమ పంట అమ్ముతున్రు.

Textbook Tamil	Madurai Spoken Tamil Dialect
இன்று பள்ளியில் விளையாட்டு போட்டி நடைபெற்றது. மாணவர்கள் ஆர்வமாக கலந்து கொண்டனர்.	இன்றுக்கு ஸ்கூல்ல விளையாட்டு போட்டி நடந்துச்சு. பசங்க ரொம்ப ஆர்வமா கலந்துக்கிட்டாங்க.

Textbook Marathi	Varhadi Marathi Dialect
काल गावात जत्रा भरली होती. लोक मोठ्या उत्साहाने सहभागी झाले.	काल गावात जत्रा भरली होती. लोक भारी उत्साहानं सहभागी झाले.

INDIAN LANGUAGE DIALECT DATA ANNOTATION CHALLENGE - 2026

THANK YOU

NEXTGENVECTORA DATA INNOVATIONS (OPC) PRIVATE LIMITED

Indian Language Dialect Data Annotation Challenge – 2026

ILDDAC–2026

NEXTGENVECTORA DATA INNOVATIONS (OPC) PRIVATE LIMITED

BUILDING THE FUTURE OF INDIAN LANGUAGE AI

PURPOSE OF THE COMPETITION

CREATE HIGH-QUALITY ANNOTATED DATASETS FOR INDIAN LANGUAGES AND DIALECTS TO SUPPORT:

🤖 NATURAL LANGUAGE PROCESSING

⚡ GENERATIVE AI MODELS

✅ RESPONSIBLE & ETHICAL AI SYSTEMS

Registration Fee
₹100 only

Join the Challenge – Register Today

Who Can Participate?

- Engineering Students
- Graduates
- Postgraduates
- Language Enthusiasts
- Researchers
- Academicians
- Developers

Three Types of Annotation Tasks

📝 *Vocabulary*
- *Word-level dialect annotation*

💭 *Sentence*
- *Phrase-level annotations*

📄 *Paragraph*
- *Context-based annotations*

https://forms.gle/VTMrCkSwdQGgTVgJ9

PRIZES & RECOGNITION

🥇 ₹25,000 – FIRST PRIZE

🥈 ₹15,000 – SECOND PRIZE

🥉 ₹10,000 – THIRD PRIZE

ADDITIONAL BENEFITS

🏆 CERTIFICATES

⭐ MERIT RECOGNITION

Important Dates
15 FEB 2026 Registration Opens
15 MAR 2026 Registration Closes
15 MAR – 15 MAY 2026 Competition Period

Open to passionate contributors in Indian language technologies