# KEC_AI_NLP@DravidianLangTech:Abusive Comment Detection in Tamil Language using Machine Learning Techniques

**Kogilavani Shanmugavadivel**[1], **Malliga Subramanian**[1], **Shri Durga R**[1], **Srigha S**[1], **Sree Harene J S**[1], **Yasvanth Bala P**[1]
[1]Department of AI, Kongu Engineering College, Perundurai, Erode.
{kogilavani.sv, mallinishanth72}@gmail.com
{shridurgar.21aim, srighas.21aim}@kongu.edu
{sreeharenijs.21aim, yasvanthbalap.21aim}@kongu.edu

## Abstract

Social media, online news reporting sites, and many other public forums on the Internet are becoming increasingly aware of abusive comments. This also leads to harassment and abusive messages that can cause anxiety and harm others. This research work aims to identify the negative comments that are associated with Counter-speech, Xenophobia, Homophobia, Transphobia, Misandry, Misogyny, and None-of-the-above categories, In order to identify these categories from the given data set, we propose three different models such as machine learning techniques, deep learning model and transfer learning model called BERT is also used to analyze the texts. In the Tamil data set, we are training the models with a train data set and testing the models with validation data. Our team participated in the shared task organized by DravidianLangTech[1] (Priyadharshini et al., 2023a)and secured fourth rank in the task of abusive comment detection in Tamil with a macro- f1 score of 0.35. Also, our run was submitted for abusive comment detection in code-mixed languages (Tamil-English) and secured sixth rank with a macro-f1 score of 0.42 using the Random Forest model.

## 1 Introduction

Social media platforms that provide communications, education, and information exchange in the digital age have become an essential part of our lives. But this increased online communication has also raised concerns about the spread of slander, hate speech, and online bullying (Park et al., 2018). Slander is a statement made against a person or group of people. In recent years there has been a growing awareness of the harm that negative comments (Narang and Brew, 2020) can do to individuals and society as a whole. Insults not only cause psychological harm to the victim, but also cause the spread of humor, division, and dissatisfaction online. Given the detrimental impact hate speech has on the general public, major platforms such as YouTube, Facebook, Instagram, and Twitter have implemented policies (Caselli et al., 2021) and protocols to address hateful content and combat negative behaviors. It is crucial to prioritize the identification and management of such comments to minimize their adverse effects on individuals.

Over the past few years, a growing body of research has been addressing the issue of tackling abusive comments (Mubarak et al., 2017) in the fields such as natural language processing (NLP), network science, and Artificial Intelligence (AI). Early studies relied on machine learning (ML) classes such as Support Vector Machine (SVM) with word and attribute n-gram features and Logistic Regression (LR) (Ibrohim and Budi, 2019). Analyzing the illegal language in Tamil requires a good understanding of the features of the language, its cultural context, and certain language patterns associated with abusive content. This has required the development of Tamil-adapted artificial intelligence algorithms and machine learning models that can identify and neutralize illegal words with higher accuracy (Davidson et al., 2019).

Machine learning models that possess specific options share similarities and present a simpler alternative to Transformer models (Koufakou et al., 2020). Recent studies on advancements in offensive language detection indicate a growing trend in utilizing deep learning-based transformer models (Mishra et al., 2019). Transfer learning is a concept where models are initially trained on extensive sets of unlabeled text using self-supervised learning, and subsequently employed on labeled text corpora. For our project, we employ the Bidirectional Encoder Representation from Transformers (BERT) model (Corazza et al., 2020).

Our task is to analyze and detect abusive com-

[1]https://codalab.lisn.upsaclay.fr/competitions/11096



*Third Workshop on Speech and Language Technologies for Dravidian Languages*,
pages 293–299, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_043

ments in Tamil (Priyadharshini et al., 2022; Shanmugavadivel et al., 2022). Tamil has an agglutinative grammar, that is, the last word is used to denote class, number case, verb tenses, and other grammatical forms. Abusive Comment Detection is a text classification problem (Priyadharshini et al., 2023b). Text classification is a technique for extracting features from the text, giving it a preset category. It is always done by the linear classifier of sentence embeddings of text. In this paper, we trained various machine learning models, deep learning, and transfer learning models for detecting abusive comments in the Tamil language. We compared the results of all the methods to determine the best model.

## 2 Literature Review

The paper by Mubarak et al. (ALW 2017) highlight the challenges associated with Arabic abusive language detection, such as the presence of dialects and informal language usage. The review in (Mubarak et al., 2017) provides valuable insights into the state-of-the-art methods, data sets, and evaluation metrics used in this domain. The paper by Mishra et al. (NAACL 2019) presents a novel approach for abusive language detection using Graph Convolutional Networks (GCNs). The authors propose a graph-based model that captures both syntactic and semantic dependencies between words in a sentence to effectively identify abusive content. (Mishra et al., 2019)leverage graph convolutional layers to learn contextual representations from the sentence's dependency graph, enabling the model to capture the structural information crucial for identifying abusive language. The paper by Park et al. (EMNLP 2018) addresses the issue of gender bias in abusive language detection systems. The authors propose a method to reduce gender bias by incorporating gender information into the training process. (Park et al., 2018) introduce a gender-balanced data set and develop a fine-tuning strategy that takes into account the gender of both the author and target of abusive language. The experimental results show that their approach effectively reduces gender bias in the detection system while maintaining good overall performance. The paper by Ibrohim and Budi (ALW 2019) focuses on the task of multi-label hate speech and abusive language detection in Indonesian Twitter. The authors present a comprehensive analysis of various approaches for addressing this challenge, including feature-based, deep learning, and ensemble methods.

(Ibrohim and Budi, 2019) discuss the importance of handling multiple labels, as hate speech and abusive language can exhibit different characteristics and require distinct detection techniques. The paper by Narang and Brew (ALW 2020) presents an approach for abusive language detection that utilizes syntactic dependency graphs. The authors propose a method that incorporates information from these graphs, including node representations and graph-based features, to identify abusive content.(Narang and Brew, 2020) demonstrate the effectiveness of this approach by comparing it with baseline models on multiple data sets and achieving superior performance. The paper by Caselli et al. (WOAH 2021) introduces HateBERT, a retraining approach for BERT (Bidirectional Encoder Representations from Transformers) specifically tailored for abusive language detection in English. The authors fine-tune BERT using a large-scale data set of annotated abusive language to enhance its ability to identify and classify abusive content accurately.

(Caselli et al., 2021) compare HateBERT's performance with other existing models and demonstrate its superiority in terms of precision, recall, and F1-score. The paper by Davidson et al. (ALW 2019) investigates the presence of racial bias in hate speech and abusive language detection data sets. The authors analyze several widely-used data sets and identify potential biases in the annotation process, particularly regarding racial and ethnic slurs. (Davidson et al., 2019) highlight the importance of addressing such biases to ensure fair and unbiased evaluation of detection models. The paper by Koufakou et al. (ALW 2020) introduces HurtBERT, a model that combines BERT with lexical features for the detection of abusive language. The authors incorporate various lexical features, such as n-grams, sentiment scores, and part-of-speech tags, alongside BERT embeddings to enhance the model's understanding of abusive content.

(Koufakou et al., 2020) evaluate HurtBERT on multiple data sets and demonstrate its improved performance compared to BERT alone and other baseline models. The paper by Corazza et al. (Findings 2020) introduces a novel approach for zero-shot abusive language detection using hybrid emoji-based masked language models. The authors propose incorporating emojis as a form of contextual information to enhance the model's ability to iden-

tify and classify abusive content. They leverage pre-trained language models and mask out abusive terms, replacing them with emoji representations during inference. The experimental results in (Corazza et al., 2020) demonstrate the effectiveness of their approach in zero-shot detection, achieving competitive performance compared to existing methods. (Chakravarthi, 2020) introduces HopeEDI, a multilingual data set for hope speech detection (Subramanian et al., 2022) in the context of equality, diversity, and inclusion. The data set aims to facilitate research on understanding and promoting positive discourse in social media. The study in (Chakravarthi, 2020) describes the data collection process, and annotation guidelines, and provides an analysis of the data set's characteristics.

## 3 Dataset Description

The goal of this shared task on abusive comment detection is to detect and reduce abusive comments on social media. The dataset used here is shared by the shared task (Priyadharshini et al., 2022). The primary goal of this project is to develop methods for detecting and classifying instances of hate speech in Tamil. The Abusive Comment Detection data set is made up of Tamil comments retrieved from the YouTube comments area (Priyadharshini et al., 2023b). The data set consists of a comment and its related label from one of the nine labels: Misandry, Counter-speech, Misogyny, Xenophobia, Hope-Speech, Homophobic, Transphobic, Not-Tamil, and None-of-the-above.SMOTE, which stands for Synthetic Minority Over-sampling data augmentation Technique, is a widely used technique in the field of machine learning specifically in the context of handling imbalanced datasets.Imbalanced datasets occur when the classes have significantly different numbers of instances, leading to a bias in the model's performance towards the majority class.

SMOTE is designed to solve this issue by generating synthetic samples for the minority class, thereby balancing the class distribution.Synthetic samples are artificially created data points that are generated based on existing data points in a dataset.In our dataset,'Misandry' belongs to the minority class,thereby we use SMOTE technique to balance the class distribution.

### 3.1 Tamil Data

The Train, Test, and Development data sets each comprise 2239, 559, 900 comments. The text in Tamil is followed by the appropriate label for each comment in the training data. In the data set, there is a significant class imbalance. Because there is no test or development data of examples for the 'Not-Tamil' label, classification is limited to the other seven labels.

| Dataset | No. of Comments |
|---|---|
| Train | 2,239 |
| Validation | 559 |
| Test | 698 |

Table 1: Dataset Description

| Class | Train | Dev | Test |
|---|---|---|---|
| None Of The Above | 1296 | 345 | 416 |
| Misandry | 446 | 104 | 127 |
| Counter Speech | 149 | 36 | 47 |
| Misogyny | 125 | 29 | 48 |
| Xenophobia | 95 | 24 | 25 |
| Hope Speech | 85 | 11 | 26 |
| Trans-phobic | 35 | 8 | 2 |

Table 2: Class Description

## 4 Methodology

Machine learning and deep learning models cannot access raw texts. Feature extraction is required to train classification models. The TF-IDF representation is utilized in ML techniques to extract features. We use three ways to analyze the results and create the best model possible: Machine Learning, Deep Learning, and Transformer-based Learning.

### 4.1 Machine Learning Models

Machine learning has come a long way in recent years, changing the way people understand important applications such as image recognition, data mining, and natural language processing(NLP). This section outlines the machine learning models utilized in the present study for text classification. We used several different kinds of machine learning algorithms such as Decision tree, Random Forest, GaussianNB, XGBoost, AdaBoost, KNN, Linear Regression, Multinomial NB, Support Vector Machine, MLP Classifier, Gradient Boost, and Ensemble models.

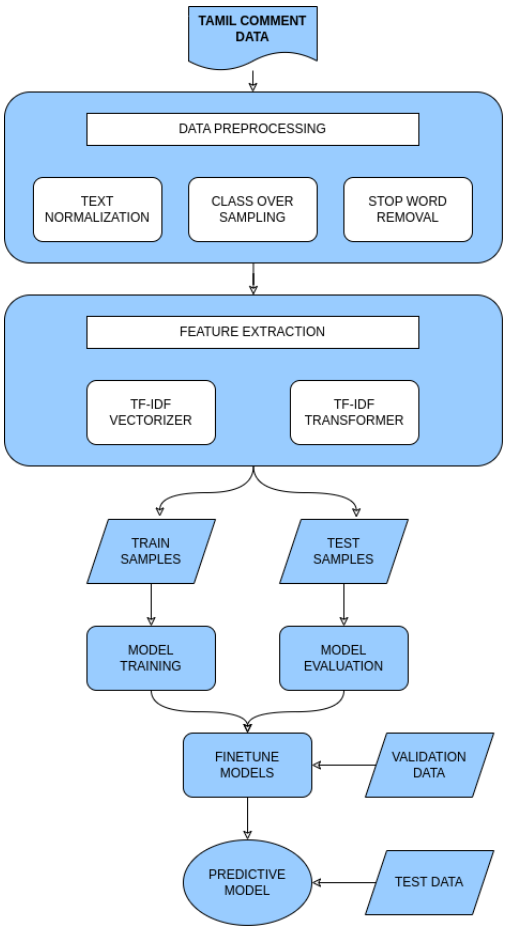


Figure 1: Proposed System Workflow

### 4.2 Deep Learning Model

Text classification tasks have witnessed the efficacy of deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. CNN models excel in capturing localized patterns and features within text data by utilizing convolutional filters. On the other hand, LSTM models are specifically designed to capture long-term dependencies and sequential information.

LSTM, which belongs to the family of recurrent neural networks (RNNs), is renowned for its capability to capture long-term dependencies in sequential data. This quality makes it highly suitable for analyzing sequences of comments. The comments are pre-processed to remove noise and irrelevant information and then fed into an LSTM model for training and evaluation.

The LSTM model is designed to learn the patterns and relationships within the comment sequences. By considering the temporal information of the comments, the model can effectively capture the context and dependencies that exist between words and phrases. The model is trained using the processed training data set, and the validation data set is used to tune the hyper-parameters and evaluate the performance of each model. Various evaluation metrics such as accuracy, precision, recall, and F1 score are used to evaluate the efficacy of the LSTM model in identifying abusive comments in Tamil.

However, it is important to note that these deep learning models can be computationally intensive and require substantial training time. This is primarily due to the high input dimensionality of text features and the large number of parameters that need to be trained. As a result, training deep learning models can often be time-consuming and resource-intensive.In conclusion, this research demonstrates the effectiveness of LSTM for abusive comment detection in YouTube content. It provides insights into the application of deep learning techniques, specifically LSTM in addressing the challenges associated with analyzing sequential comment data.

### 4.3 Transfer Learning Model

Transfer learning is a powerful technique that leverages pre-trained models to enhance the performance of models on new tasks. BERT(Bidirectional Encoder Representations from Transformers) is a powerful language model that has achieved remarkable results in various natural language processing tasks. To leverage the capabilities of BERT, it is possible to utilize pre-trained BERT models that have been trained on extensive amounts of text data from diverse sources.

The pre-trained BERT model is tuned-up on the particular task of abusive comment detection using the YouTube comments data set. Fine-tuning involves adapting the previously trained model to the target task by training it on the labeled data. This process allows BERT to learn the specific patterns and features relevant to identifying abusive comments in YouTube content.

During the fine-tuning process, appropriate classification techniques, such as adding a classification layer on top of BERT or employing additional neural network layers for improved performance can be done. The fine-tuned BERT model is evaluated on a separate test set to measure its accuracy score, precision, recall, and F1 score metrics. Additionally,the advantages of transfer learning with BERT include the ability to capture semantic meaning, context, and nuanced language patterns. It also addresses potential limitations, such as the need for large amounts of labeled data and computational resources for fine-tuning BERT.

## 5 Performance Evaluation

From the comprehensive results presented in Table 3 and Table 4, it becomes apparent that out of the 13 models tested, comprising 11 machine learning models, deep learning model, and transfer learning model, SVC, a machine learning model, emerged as the top performer in term of precision and recall. It outperformed both the deep learning and transfer learning models in terms of precision, recall, and F1 score, signifying its superiority in predictive capabilities. Figure 2 displayed the confusion matrix of the Random Forest Model.

With its ability to leverage an ensemble of decision trees and feature importance estimation, the random forest model demonstrated its prowess in capturing complex patterns within the data set and making accurate predictions. The performance advantage of the random forest model can be attributed to its ability to handle high-dimensional data, effectively deal with noisy and missing values, and mitigate overfitting concerns. By leveraging a combination of feature subsampling and bootstrap aggregating, the model achieved robust generalization and reduced the risk of overfitting.

The deep learning model, which often requires significant computational resources and extensive parameter tuning, fell short in this evaluation. It might have struggled to extract meaningful representations from the given data set or faced challenges in optimizing its numerous parameters, thereby resulting in comparatively lower accuracy and F1 score. Similarly, the transfer learning model, which typically leverages pre-trained neural networks and fine-tunes them for specific tasks, failed to outperform the random forest model. Although transfer learning has proven to be effective in various domains, it appears that the unique characteristics of the data set did not lend themselves well to this particular transfer learning approach.

Overall, the exceptional performance of the support vector classifier model highlights the strength of traditional machine learning algorithms, particularly in scenarios where the data-set is not excessively large or lacks the complexity that deep learning models excel at handling. This outcome underscores the importance of carefully selecting the appropriate modeling technique based on the specific characteristics and requirements of the problem at hand, ultimately leading to improved predictive performance.

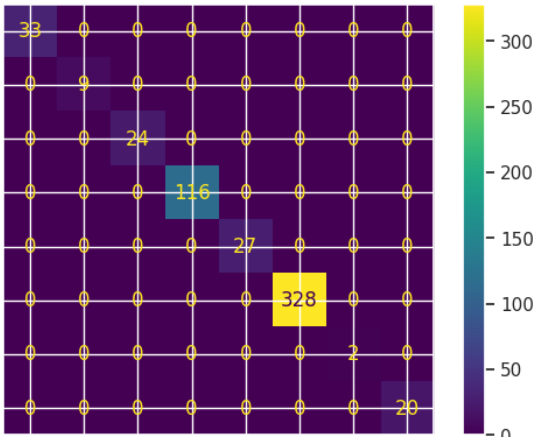


Figure 2: Confusion Matrix Of Support Vector Classifier Model- Train Data

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Multilayer perceptron | 0.80 | 0.66 | 0.70 |
| K-Nearest Neighbour | 0.66 | 0.35 | 0.34 |
| Xtreme Gradient Boost | 0.81 | 0.69 | 0.70 |
| Decision Tree | 0.80 | 0.79 | 0.79 |
| Random Forest | 0.85 | 0.81 | 0.81 |
| Logistic Regression | 0.81 | 0.83 | 0.85 |
| Support Vector Classifier | 0.99 | 0.99 | 0.99 |
| Multinomial Naive Bayes | 0.60 | 0.53 | 0.64 |
| Gradient Boost Classifier | 0.94 | 0.94 | 0.94 |
| Ensemble | 0.98 | 0.98 | 0.98 |
| Adaboost Classifier | 0.52 | 0.55 | 0.64 |
| BERT | 0.76 | 0.76 | 0.79 |
| CNN | 0.65 | 0.52 | 0.62 |

Table 3: Tamil-Validation-Train Data Evaluation Metrics

| Model | Precision | Recall | F1-score |
|---|---|---|---|
| Multilayer perceptron | 0.45 | 0.45 | 0.54 |
| K-Nearest Neighbour | 0.51 | 0.51 | 0.44 |
| Xtreme Gradient Boost | 0.55 | 0.52 | 0.46 |
| Decision Tree | 0.46 | 0.52 | 0.54 |
| Random Forest | 0.49 | 0.52 | 0.44 |
| Logistic Regression | 0.66 | 0.68 | 0.62 |
| Support Vector Classifier | 0.64 | 0.67 | 0.64 |
| Multinomial Naive Bayes | 0.53 | 0.60 | 0.47 |
| Gradient Boost Classifier | 0.65 | 0.68 | 0.65 |
| Ensemble | 0.66 | 0.69 | 0.64 |
| Adaboost Classifier | 0.64 | 0.64 | 0.55 |
| BERT | 0.59 | 0.66 | 0.60 |
| CNN | 0.41 | 0.50 | 0.45 |

Table 4: Tamil-Train Data Evaluation Metrics

## 6 Conclusion

The study focuses on the detection of abusive comments in the Tamil language. We compare the performance of different models in this task. Deep Learning and Transformer models did not achieve superior results when trained and evaluated on Tamil data. Instead, Machine Learning models outperformed the Deep Learning and Transformer-based models. It is important to note that contextualized embeddings such as ELMO or FLAIR, which have shown potential in enhancing the performance of language models has not been utilized. The absence of these embeddings might have limited the effectiveness of the models used in our study.

We acknowledge this limitation and suggest that future work should explore the implementation of contextualized embeddings using deep learning techniques. We believe that incorporating these advanced embeddings could potentially improve the detection of abusive comments in Tamil. Additionally, other models like Indic BERT and Muril BERT were not utilized at this stage. However, we highlight the possibility of implementing these models in the future as our project advances. This indicates the potential for further exploration and improvement in the detection of abusive comments in Tamil using more advanced language models.

## References

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. Dalc: the dutch abusive language corpus. In *Proceedings of the 5th*

*Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66.

Bharathi Raja Chakravarthi. 2020. Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 41–53.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Muhammad Okky Ibrohim and Indra Budi. 2019. Multi-label hate speech and abusive language detection in indonesian twitter. In *Proceedings of the third workshop on abusive language online*, pages 46–57.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, Viviana Patti, et al. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the fourth workshop on online abuse and harms*, pages 34–43. Association for Computational Linguistics.

Pushkar Mishra, Marco Del Tredici, Helen Yannakoudakis, and Ekaterina Shutova. 2019. Abusive language detection with graph convolutional networks. *arXiv preprint arXiv:1904.04073*.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on arabic social media. In *Proceedings of the first workshop on abusive language online*, pages 52–56.

Kanika Narang and Chris Brew. 2020. Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 44–53.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in Tamil-ACL 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, SUBALALITHA CN, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023a. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga Subramanian, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Premjith B, Abirami Murugappan, Prasanna Kumar Kumaresan, Karnati Sai Prashanth, Mangamuru Sai Rishith Reddy, and Janakiram Chandu. 2023b. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. *DravidianLangTech*, 2022:292.

Malliga Subramanian, Ramya Chinnasamy, Prasanna Kumar Kumaresan, Vasanth Palanikumar, Madhoora Mohan, and Kogilavani Shanmugavadivel. 2022. Development of multi-lingual models for detecting hope speech texts from social media comments. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 209–219. Springer.