Learning to Recognize Dialect Features

Dorottya Demszky[1]* **Devyani Sharma**[2] **Jonathan H. Clark**[3]
Vinodkumar Prabhakaran[3] **Jacob Eisenstein**[3]
[1]Stanford Linguistics [2]Queen Mary University of London [3]Google Research
ddemszky@stanford.edu
d.sharma@qmul.ac.uk
{jhclark,vinodkpg,jeisenstein}@google.com

Abstract

Building NLP systems that serve everyone requires accounting for dialect differences. But dialects are not monolithic entities: rather, distinctions between and within dialects are captured by the presence, absence, and frequency of dozens of dialect features in speech and text, such as the deletion of the copula in "He ∅ running". In this paper, we introduce the task of dialect feature detection, and present two multitask learning approaches, both based on pretrained transformers. For most dialects, large-scale annotated corpora for these features are unavailable, making it difficult to train recognizers. We train our models on a small number of minimal pairs, building on how linguists typically define dialect features. Evaluation on a test set of 22 dialect features of Indian English demonstrates that these models learn to recognize many features with high accuracy, and that a few minimal pairs can be as effective for training as thousands of labeled examples. We also demonstrate the downstream applicability of dialect feature detection both as a measure of dialect density and as a dialect classifier.

1 Introduction

Dialect variation is a pervasive property of language, which must be accounted for if we are to build robust natural language processing (NLP) systems that serve everyone. Linguists do not characterize dialects as simple categories, but rather as collections of correlated features (Nerbonne, 2009), such as the one shown in Figure 1; speakers of any given dialect vary regarding which features they employ, how frequently, and in which contexts. In comparison to approaches that classify speakers or documents across dialects (typically using metadata such as geolocation), the feature-based perspective has several advantages: (1) allowing for fine-grained comparisons of speakers or documents

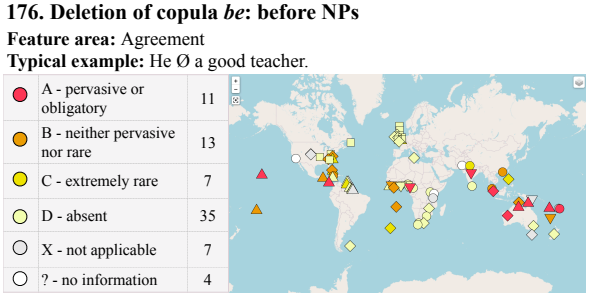

Figure 1: An example dialect feature from the Electronic World Atlas of Varieties of English (eWAVE).[1]

within dialects, without training on personal metadata; (2) disentangling grammatical constructions that make up the dialect from the content that may be frequently discussed in the dialect; (3) enabling robustness testing of NLP systems across dialect features, helping to ensure adequate performance even on cases other than "high-resource" varieties such as mainstream U.S. English (Blodgett et al., 2016); (4) helping to develop more precise characterizations of dialects, enabling more accurate predictions of variable language use and better interpretations of its social implications (e.g., Craig and Washington, 2002; Van Hofwegen and Wolfram, 2010).

The main challenge for recognizing dialect features computationally is the lack of labeled data. Annotating dialect features requires linguistic expertise and is prohibitively time-consuming given the large number of features and their sparsity. In dialectology, large-scale studies of text are limited to features that can be detected using regular expressions of surface forms and parts-of-speech, e.g., `PRP DT` for the copula deletion feature in Figure 1; many features cannot be detected with such patterns (e.g. OBJECT FRONTING, EXTRANEOUS ARTICLE). Furthermore, part-of-speech tagging is unreliable in many language varieties, such as re-

* Work done while at Google Research.

[1] https://ewave-atlas.org. Shapes indicate variety type, e.g. creole, L1, and L2 English varieties.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2315–2338
June 6–11, 2021.

gional and minority dialects (Jørgensen et al., 2015; Blodgett et al., 2016). As dialect density correlates with social class and economic status (Sahgal and Agnihotri, 1988; Rickford et al., 2015; Grogger et al., 2020), the failure of language technology to cope with dialect differences may create allocational harms that reinforce social hierarchies (Blodgett et al., 2020).

In this paper, we propose and evaluate learning-based approaches to recognize dialect features. We focus on Indian English, given the availability of domain expertise and labeled corpora for evaluation. First, we consider a standard multitask classification approach, in which a pretrained transformer (Vaswani et al., 2017) is fine-tuned to recognize a set of dialect features. The architecture can be trained from two possible sources of supervision: (1) thousands of labeled corpus examples, (2) a small set of *minimal pairs*, which are hand-crafted examples designed to highlight the key aspects of each dialect feature (as in the "typical example" field of Figure 1). Because most dialects have little or no labeled data, the latter scenario is more realistic for most dialects. We also consider a multitask architecture that learns across multiple features by encoding the feature names, similar to recent work on few-shot or zero-shot multitask learning (Logeswaran et al., 2019; Brown et al., 2020).

In Sections 4 and 5, we discuss empirical evaluations of these models. Our main findings are:

- It is possible to detect individual dialect features: several features can be recognized with reasonably high accuracy. Our best models achieve a macro-AUC of .848 across ten grammatical features for which a large test set is available.
- This performance can be obtained by training on roughly five minimal pairs per feature. Minimal pairs are significantly more effective for training than a comparable number of corpus examples.
- Dialect feature recognizers can be used to rank documents by their density of dialect features, enabling within-dialect density computation for Indian English and accurate classification between Indian and U.S. English.

2 Data and Features of Indian English

We develop methods for detecting 22 dialect features associated with Indian English. Although India has over 125 million English speakers — making it the world's second largest English-speaking population — there is relatively little NLP research focused on Indian English. Our methods are not designed exclusively for specific properties of Indian English; many of the features that are associated with Indian English are also present in other dialects of English.

We use two sources of data in our study: an annotated corpus (§ 2.1) and a dataset of minimal pairs (§ 2.2). For evaluation, we use corpus annotations exclusively. The features are described in Table 1, and our data is summarized in Table 2.

2.1 Corpus Annotations

The International Corpus of English (ICE; Greenbaum and Nelson, 1996) is a collection of corpora of world varieties of English, organized primarily by the national origin of the speakers/writers. We focus on annotations of spoken dialogs (S1A-001 – S1A-090) from the Indian English subcorpus (ICE-India). The ICE-India subcorpus was chosen in part because it is one of the only corpora with large-scale annotations of dialect features. To contrast Indian English with U.S. English (§ 4), we use the Santa Barbara Corpus of Spoken American English (Du Bois et al., 2000) that constitutes the ICE-USA subcorpus of spoken dialogs.

We work with two main sources of dialect feature annotations in the ICE-India corpus:

Lange features. The first set of annotations come from Claudia Lange (2012), who annotated 10 features in 100 transcripts for an analysis of discourse-driven syntax in Indian English, such as topic marking and fronting. We use half of this data for training (50 transcripts, 9392 utterances), and half for testing (50 transcripts, 9667 utterances).

Extended features. To test a more diverse set of features, we additionally annotated 18 features on a set of 300 turns randomly selected from the conversational subcorpus of ICE-India,[2] as well as 50 examples randomly selected from a secondary dataset of sociolinguistic interviews (Sharma, 2009) to ensure diverse feature instantiation. We selected our 18 features based on multiple criteria: 1) prevalence in Indian English based on the dialectology literature, 2) coverage in the data (we started out with a larger set of features and removed those with fewer than two occurrences), 3) diversity of linguistic phenomena. The extended

[2]We manually split turns that were longer than two clauses, resulting in 317 examples.

Feature	Example	Count of Instantiations	
		Lange (2012)	Our data
ARTICLE OMISSION	*(the) chair is black*		59
DIRECT OBJECT PRO-DROP	*she doesn't like (it)*		14
FOCUS *itself*	*he is doing engineering in Delhi itself*	24	5
FOCUS *only*	*I was there yesterday only*	95	8
HABITUAL PROGRESSIVE	*always we are giving receipt*		2
STATIVE PROGRESSIVE	*he is having a television*		3
LACK OF INVERSION IN WH-QUESTIONS	*what you are doing?*		4
LACK OF AGREEMENT	*he do a lot of things*		23
LEFT DISLOCATION	*my father, he works for a solar company*	300	19
MASS NOUNS AS COUNT NOUNS	*all the musics are very good*		13
NON-INITIAL EXISTENTIAL	*every year inflation is there*	302	8
OBJECT FRONTING	*minimum one month you have to wait*	186	14
PP FRONTING WITH REDUCTION	*(on the) right side we can see a plate*		11
PREPOSITION OMISSION	*I went (to) another school*		17
INVERSION IN EMBEDDED CLAUSE	*I don't know what are they doing*		4
INVARIANT TAG *(isn't it, no, na)*	*the children are outside, isn't it?*	786	17
EXTRANEOUS ARTICLE	*she has a business experience*		25
GENERAL EXTENDER *and all*	*then she did her schooling and all*		7
COPULA OMISSION	*my parents (are) from Gujarat*	71	
RESUMPTIVE OBJECT PRONOUN	*my old life I want to spend it in India*	24	
RESUMPTIVE SUBJECT PRONOUN	*my brother, he lives in California*	287	
TOPICALIZED NON-ARGUMENT CONSTITUENT	*in those years I did not travel*	272	

Table 1: Features of Indian English used in our evaluations and their counts in the two datasets we study.

Dialect features		Unique annotated examples	
Feature set	Count	Corpus ex.	Min. pair ex.
Lange (2012)	10	19059	113
Extended	18	367	208

Table 2: Summary of our labeled data. All corpus examples for the Lange features are from ICE-India; for the Extended feature set, examples are drawn from ICE-India and the Sharma data.

features overlap with those annotated by Lange, yielding a total set of 22 features. Annotations were produced by consensus from the first two authors. To measure interrater agreement, a third author (JE) independently re-annotated 10% of the examples, with Cohen's $\kappa = 0.79$ (Cohen, 1960).[3]

2.2 Minimal Pairs

For each of the 22 features in Table 1, we created a small set of minimal pairs. The pairs were created by first designing a short example that demonstrated the feature, and then manipulating the example so that the feature is absent. This "negative" example captures the *envelope of variation* for the feature, demonstrating a site at which the feature could be applied (Labov, 1972). Consequently, negative examples in minimal pairs carry more information than in the typical annotation scenario, where absence of evidence does not usually imply evidence of absence. In our minimal pairs, the negative examples were chosen to be acceptable in standard U.S. and U.K. English, and can thus be viewed as situating dialects against standard varieties. Here are some example minimal pairs:

ARTICLE OMISSION: *chair is black* → *the chair is black*

FOCUS *only*: *I was there yesterday only* → *I was there just yesterday.*

NON-INITIAL EXISTENTIAL: *every year inflation is there* → *every year there is inflation.*

For most features, each minimal pair contains exactly one positive and one negative example. However, in some cases where more than two variants are available for an example (e.g., for the feature INVARIANT TAG *(isn't it, no, na)*), we provide multiple positive examples to illustrate different variants. For Lange's set of 10 features, we provide a total of 113 unique examples; for the 18 extended features, we provide a set of 208 unique examples, roughly split equally between positives and negatives. The complete list of minimal pairs is included in Appendix D.

[3]Our annotations will be made available at https://dialectfeatures.page.link/annotations.

y	x
1	[CLS] article omission [SEP] Chair is black. [SEP]
0	[CLS] article omission [SEP] The chair is black. [SEP]
0	[CLS] article omission [SEP] I was there yesterday only. [SEP]
...	...
1	[CLS] focus only [SEP] I was there yesterday only. [SEP]
0	[CLS] focus only [SEP] I was there just yesterday. [SEP]
0	[CLS] focus only [SEP] Chair is black. [SEP]
...	...

Figure 2: Conversion of minimal pairs to labeled examples for DAMTL, using two minimal pairs.

3 Models and training

We train models to recognize dialect features by fine-tuning the BERT-base uncased transformer architecture (Devlin et al., 2019). We consider two strategies for constructing training data, and two architectures for learning across multiple features.

3.1 Sources of supervision

We consider two possible sources of supervision:

Minimal pairs. We apply a simple procedure to convert minimal pairs into training data for classification. The positive part of each pair is treated as a positive instance for the associated feature, and the negative part is treated as a negative instance. Then, to generate more data, we also include elements of other minimal pairs as examples for each feature: for instance, a positive example of the RESUMPTIVE OBJECT PRONOUN feature would be a negative example for FOCUS *only*, unless the example happened to contain both features (this was checked manually). In this way, we convert the minimal pairs into roughly 113 examples per feature for Lange's features and roughly 208 examples per feature for the extended features. The total number of unique surface forms is still 113 and 208 respectively. Given the lack of labeled data for most dialects of the world, having existing minimal pairs or collecting a small number of minimal pairs is the most realistic data scenario.

Corpus annotations. When sufficiently dense annotations are available, we can train a classifier based on these labeled instances. We use 50 of the ICE-India transcripts annotated by Lange, which consists of 9392 labeled examples (utterances) per feature. While we are lucky to have such a large resource for the Indian English dialect, this high-resource data scenario is rare.

3.2 Architectures

We consider two classification architectures:

Multihead. In this architecture, which is standard for multitask classification, we estimate a linear *prediction head* for each feature, which is simply a vector of weights. This is a multitask architecture, because the vast majority of model parameters from the input through the deep BERT stack remain shared among dialect features. The prediction head is then multiplied by the BERT embedding for the [CLS] token to obtain a score for a feature's applicability to a given instance.

DAMTL. Due to the few-shot nature of our prediction task, we also consider an architecture that attempts to exploit the natural language descriptions of each feature. This is done by concatenating the feature description to each element of the minimal pair. The instance is then labeled for whether the feature is present. This construction is shown in Figure 2. Prediction is performed by learning a single linear prediction head on the [CLS] token. We call this model *description-aware multitask learning*, or DAMTL.

Model details. Both architectures are built on top of the BERT-base uncased model, which we fine-tune by cross-entropy for 500 epochs (due to the small size of the training data) using the Adam optimizer (Kingma and Ba, 2014), batch size of 32 and a learning rate of 10^{-5}, warmed up over the first 150 epochs. Annotations of dialect features were not used for hyperparameter selection. Instead, the hyperparameters were selected to maximize the discriminability between corpora of Indian and U.S. English, as described in § 5.2. All models trained in less than two hours on a pod of four v2 TPU chips, with the exception of DAMTL on corpus examples, which required up to 18 hours.

3.3 Regular Expressions

In dialectology, regular expression pattern matching is the standard tool for recognizing dialect features (e.g., Nerbonne et al., 2011). For the features

Supervision:	Corpus examples		Minimal pairs	
Dialect feature	DAMTL	Multihead	DAMTL	Multihead
FOCUS *itself**	0.945	0.925	0.974	0.960
FOCUS *only**	0.975	0.911	0.994	0.938
INVARIANT TAG	0.991	0.985	0.969	0.925
COPULA OMISSION	0.536	0.641	0.626	0.746
LEFT DISLOCATION	0.855	0.879	0.765	0.885
NON-INITIAL EXISTENTIAL*	0.991	0.992	0.905	0.879
OBJECT FRONTING	0.805	0.809	0.678	0.761
RES. OBJECT PRONOUN	0.595	0.667	0.733	0.825
RES. SUBJECT PRONOUN	0.886	0.887	0.688	0.857
TOPICALIZED NON-ARG. CONST.	0.725	0.727	0.499	0.707
Macro Average	0.830	0.842	0.783	0.848

Table 3: ROC-AUC scores on the Lange feature set, averaged across five random seeds. Asterisk (*) marks features that can be detected with relatively high accuracy (> 0.85 ROC-AUC) using regular expressions.

described in Table 1, we were able to design regular expressions for only five.[4] Prior work sometimes relies on patterns that include both surface forms and part-of-speech (e.g., Bohmann, 2019), but part-of-speech cannot necessarily be labeled automatically for non-standard dialects (Jørgensen et al., 2015; Blodgett et al., 2016), so we consider only regular expressions over surface forms.

4 Results on Dialect Feature Detection

In this section, we present results on the detection of individual dialect features. Using the features shown in Table 1, we compare supervision sources (corpus examples versus minimal pairs) and classification architectures (multihead versus DAMTL) as described in § 3. To avoid tuning a threshold for detection, we report area under the ROC curve (ROC-AUC), which has a value of .5 for random guessing and 1 for perfect prediction.[5]

4.1 Results on Lange Data and Features

We first consider the 10 syntactic features from Lange (2012), for which we have large-scale annotated data: the 100 annotated transcripts from the ICE-India corpus are split 50/50 into training and test sets. As shown in Table 3, it is possible to achieve a Macro-AUC approaching .85 overall with multihead predictions on minimal pair examples. This is promising, because it suggests the possibility of recognizing dialect features for which we lack labeled corpus examples – and such low-data situations are by far the most common data scenario among the dialects of the world.

The multihead architecture outperforms DAMTL on both corpus examples and minimal pairs. In an ablation, we replaced the feature descriptions with non-descriptive identifiers such as "Feature 3". This reduced the Macro-AUC from to .80 with corpus examples, and to .76 with minimal pairs (averaged over five random seeds). We also tried longer feature descriptions, but this did not improve performance.

Unsurprisingly, the lexical features (e.g., FOCUS *itself*) are easiest to recognize. The more syntactical features (e.g., COPULA OMISSION, RESUMPTIVE OBJECT PRONOUN) are more difficult, although some movement-based features (e.g., LEFT DISLOCATION, RESUMPTIVE SUBJECT PRONOUN) can be recognized accurately.

Qualitative model comparison. We conducted a qualitative comparison of three models: regular expressions and two versions of the multihead model, one trained on corpus examples and another trained on minimal pairs. Table 4 includes illustrative examples for the Lange data and features where models make different predictions. We find that the minimal pair model is better able to account for rare cases (e.g. use of non-focus "only" in Example 1), likely as it was trained on a few carefully selected set of examples illustrating positives and negatives. Both multihead models are able to account for disfluencies and restarts, in contrast to regular expressions (Example 2). Our analysis shows that several model errors are accounted for by difficult examples (Example 3: "is there" followed by "isn't"; Example 6: restart mistaken for left dislocation) or the lack of contextual information available to the model (Example 4 & 7: truncated examples). Please see Appendix B for more details and random samples of model predictions.

Learning from fewer corpus examples. The minimal pair annotations consist of 113 examples; in contrast, there are 9392 labeled corpus examples, requiring far more effort to create. We now consider the situation when the amount of labeled data is reduced, focusing on the Lange features (for which labeled training data is available). As shown in Figure 3, even 5000 labeled corpus examples do not match the performance of training on roughly 5 minimal pairs per feature.

[4] Features: FOCUS *itself*, FOCUS *only*, NON-INITIAL EXISTENTIAL, INVARIANT TAG *(isn't it, no, na)*, and GENERAL EXTENDER *and all*. Table 7 lists all regular expressions.

[5] Results for area under the precision-recall (AUPR) curve are shown in Appendix C. According to this metric, minimal pairs are less effective than the full training set of corpus examples, on average.

	Example	Feature	Gold label	Regex	Multihead	
					Corpus ex.	Min. pair
1	But whereas in Hyderabad they are only stuck with their books and home and work that's all like	FOCUS *only*	0	1	1	0
2	There is there is a club this humour club oh good and I've chance I had a chance of attending	NON-INITIAL EXISTENTIAL	0	1	0	0
3	New Education Policy is there isn't it?	NON-INITIAL EXISTENTIAL	1	1	0	0
4	I didn't go anywhere no	INVARIANT TAG *(isn't it, no, na)*	0	1	1	1
5	In fact my son and daughter they had asked me to buy buy them this thing the sunglasses	LEFT DISLOCATION	1	N/A	1	1
6	His house he is going to college KK diploma electronics	RESUMPTIVE SUBJECT PRONOUN	0	N/A	0	1
7	Which October first I think	COPULA OMISSION	0	N/A	1	0
8	Papers we can't say hard only because they already taught that same	COPULA OMISSION	1	N/A	0	0
9	Just typing work I have to do	OBJECT FRONTING	1	N/A	1	1
10	My post graduation degree I finished it in mid June nineteen eighty-six	RESUMPTIVE OBJECT PRONOUN	1	N/A	0	1

Table 4: Example model predictions from the Lange data and feature set, comparing regular expressions with two versions of the multihead model, one trained on corpus examples and another on minimal pairs. 'Gold label' indicates whether the feature was manually labeled as present in the original Lange data. Green and red indicate correct and incorrect predictions, respectively.

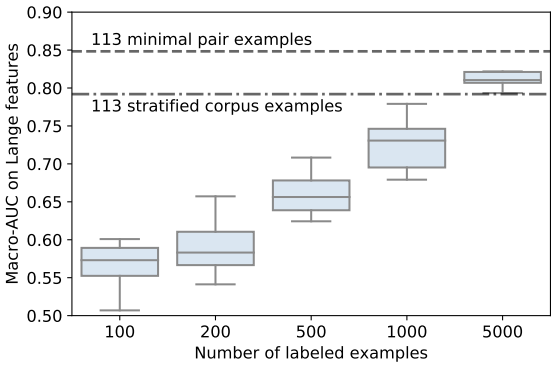

Figure 3: Performance of the multihead model as the number of corpus examples is varied. Box plots are over 10 random data subsets, showing the 25th, 50th, and 75th percentiles; whiskers show the most extreme points within ± 1.5 times the inter-quartile range.

Corpus examples stratified by feature. One reason that subsampled datasets yield weaker results is that they lack examples for many features. To enable a more direct comparison of corpus examples and minimal pairs, we created a set of "stratified" datasets of corpus examples, such that the number of positive and negative examples for each feature exactly matches the minimal pair data. Averaged over ten such random stratified samples, the multihead model achieves a Macro-AUC of .790 ($\sigma = 0.029$), and DAMTL achieves a Macro-AUC of .722 ($\sigma = .020$). These results are considerably worse than training on an equivalent number of minimal pairs, where the multihead model achieves a Macro-AUC of .848 and DAMTL achieves a Macro-AUC of .783. This demonstrates the utility of minimal pairs over corpus examples for learning to recognize dialect features.

4.2 Results on Extended Feature Set

Next, we consider the extended features, for which we have sufficient annotations for testing but not training (Table 1). Here we compare the DAMTL and multihead models, using minimal pair data in both cases. As shown in Table 5, performance on these features is somewhat lower than on the Lange features, and for several features, at least one of the recognizers does worse than chance: DIRECT OBJECT PRO-DROP, EXTRANEOUS ARTICLE, MASS NOUNS AS COUNT NOUNS. These features seem to require deeper syntactic and semantic analysis, which may be difficult to learn from a small number of minimal pairs. On the other extreme, features with a strong lexical signature are recognized with high accuracy: GENERAL EXTENDER *and all*, FOCUS *itself*, FOCUS *only*. These three features can also be recognized by regular expressions, as can NON-INITIAL EXISTENTIAL.[6] However, for a number of other features, it is possible to learn a fairly accurate recognizer from just five minimal pairs: ARTICLE OMISSION, INVERSION IN EMBEDDED CLAUSE, LEFT DISLOCATION, LACK OF INVERSION IN WH-QUESTIONS.

[6] `\band all\b, \bitself\b, \bonly\b, \bis there\b|\bare there\b`

Dialect feature	DAMTL	Multihead
ARTICLE OMISSION	0.581	0.658
DIRECT OBJECT PRO-DROP	0.493	0.563
EXTRANEOUS ARTICLE	0.546	0.465
FOCUS *itself**	1.000	0.949
FOCUS *only**	0.998	0.775
HABITUAL PROGRESSIVE	0.439	0.718
INVARIANT TAG	0.984	0.901
INVERSION IN EMBEDDED CLAUSE	0.719	0.884
LACK OF AGREEMENT	0.543	0.674
LACK OF INVERSION IN WH-QUESTIONS	0.649	0.660
LEFT DISLOCATION	0.758	0.820
MASS NOUNS AS COUNT NOUNS	0.443	0.465
NON-INITIAL EXISTENTIAL*	0.897	0.885
OBJECT FRONTING	0.722	0.789
PREPOSITION OMISSION	0.500	0.648
PP FRONTING WITH REDUCTION	0.655	0.697
STATIVE PROGRESSIVE	0.645	0.789
GENERAL EXTENDER *and all*	0.994	0.991
Macro Average	0.698	0.741

Table 5: ROC-AUC results on the extended feature set, averaged across five random seeds. Because labeled corpus examples are not available for some features, we train only on minimal pairs. Asterisk (*) marks features that can be detected with relatively high accuracy ($>$ 0.85 ROC-AUC) using regular expressions.

4.3 Summary of Dialect Feature Detection

Many dialect features can be automatically recognized with reasonably high discriminative power, as measured by area under the ROC curve. However, there are also features that are difficult to recognize: particularly, features of omission (such as DIRECT OBJECT PRO-DROP and PREPOSITION OMISSION), and the more semantic features such as MASS NOUNS AS COUNT NOUNS. While some features can also be identified through regular expressions (e.g., FOCUS *only*), there are many features that can be learned but cannot be recognized by regular expressions. We now move from individual features to aggregate measures of dialect density.

5 Measuring Dialect Density

A dialect density measure (DDM) is an aggregate over multiple dialect features that tracks the vernacularity of a passage of speech or text. Such measures are frequently used in dialectology (Van Hofwegen and Wolfram, 2010), and are also useful in research on education (e.g., Craig and Washington, 2002). Recently, a DDM was used to evaluate the performance of speech recognition systems by the density of AAVE features (Koenecke et al., 2020). The use of DDMs reflects the reality that speakers construct individual styles drawing on linguistic repertoires such as dialects to varying degrees (Benor, 2010). This necessitates a more nuanced description for speakers and texts than a discrete dialect category.

Following prior work (e.g., Van Hofwegen and Wolfram, 2010) we construct dialect density measures from feature detectors by counting the predicted number of features in each utterance, and dividing by the number of tokens. For the learning-based feature detectors (minimal pairs and corpus examples), we include partial counts from the detection probability; for the regular expression detectors, we simply count the number of matches and dividing by the number of tokens. In addition, we construct a DDM based on a document classifier: we train a classifier to distinguish Indian English from U.S. English, and then use its predictive probability as the DDM. These DDMs are then compared on two tasks: distinguishing Indian and U.S. English, and correlation with the density of expert-annotated features. The classifier is trained by fine-tuning BERT, using a prediction head on the [CLS] token.

5.1 Ranking documents by dialect density

One application of dialect feature recognizers is to rank documents based on their dialect density, e.g. to identify challenging cases for evaluating downstream NLP systems, or for dialectology research. We correlate the dialect density against the density of expert-annotated features from Lange (2012), both measured at the transcript-level, and report the Spearman rank-correlation ρ.

As shown in Table 6, the document classifier performs poorly: learning to distinguish Indian and U.S. English offers no information on the density of Indian dialect features, suggesting that the model is attending to other information, such as topics or entities. The feature-based model trained on labeled examples performs best, which is unsurprising because it is trained on the same type of features that it is now asked to predict. Performance is weaker when the model is trained from minimal pairs. Minimal pair training is particularly helpful on rare features, but offers far fewer examples on the high-frequency features, which in turn dominate the DDM scores on test data. Regular expressions perform well on this task, because we happen to have regular expressions for the high-frequency features, and because the precision issues are less problematic in aggregate when the DDM is not applied to non-dialectal transcripts.

5.2 Dialect Classification

Another application of dialect feature recognizers is to classify documents or passages by dialect (Dunn, 2018). This can help to test the performance of downstream models across dialects, assessing dialect transfer loss (e.g., Blodgett et al., 2016), as well as identifying data of interest for manual dialectological research. We formulate a classification problem using the ICE-India and the Santa Barbara Corpus (ICE-USA). Each corpus is divided into equal-size training and test sets. The training corpus was also used for hyperparameter selection for the dialect feature recognition models, as described in § 3.2.

The dialect classifier was constructed by building on the components from § 5.1. For the test set, we measure the D' ("D-prime") statistic (Macmillan and Creelman, 1991),

$$D' = \frac{\mu_{\text{IN}} - \mu_{\text{US}}}{\sqrt{\frac{1}{2}(\sigma^2_{\text{IN}} + \sigma^2_{\text{US}})}}. \quad (1)$$

This statistic, which can be interpreted similarly to a Z-score, quantifies the extent to which a metric distinguishes between the two populations. We also report classification accuracy; lacking a clear way to set a threshold, for each classifier we balance the number of false positives and false negatives.

As shown in Table 6, both the document classifier and the corpus-based feature detection model (trained on labeled examples) achieve high accuracy at discriminating U.S. and Indian English. The D' discriminability score is higher for the document classifier, which is trained on a cross-entropy objective that encourages making confident predictions. Regular expressions suffer from low precision because they respond to surface cues that may be present in U.S. English, even when the dialect feature is not present (e.g., the word "only", the phrase "is there").

	Ranking	Classification	
Dialect density measure	ρ	D'	acc.
Document classifier	−0.17	14.48	1
Multihead, corpus examples	0.83	2.30	0.95
Multihead, minimal pairs	0.70	1.85	0.85
Regular expressions	0.71	1.61	0.80

Table 6: Performance of dialect density measures at the tasks of ranking Indian English transcripts by dialect density (quantified by Spearman ρ) and distinguishing Indian and U.S. English transcripts (quantified by accuracy and D' discriminability).

6 Related Work

Dialect classification. Prior work on dialect in natural language processing has focused on distinguishing between dialects (and closely-related languages). For example, the VarDial 2014 shared task required systems to distinguish between nation-level language varieties, such as British versus U.S. English, as well as closely-related language pairs such as Indonesian versus Malay (Zampieri et al., 2014); later evaluation campaigns expanded this set to other varieties (Zampieri et al., 2017). In general, participants in these shared tasks have taken a text classification approach; neural architectures have appeared in the more recent editions of these shared tasks, but with a few exceptions (e.g., Bernier-Colborne et al., 2019), they have not outperformed classical techniques such as support vector machines. Our work differs by focusing on a specific set of known dialect features, rather than document-level classification between dialects, which aligns with the linguistic view of dialects as bundles of correlated features (Nerbonne, 2009) and tracks variable realization of features within dialect usage.

Discovering and detecting dialect features. Machine learning feature selection techniques have been employed to discover dialect features from corpora. For example, Dunn (2018, 2019) induces a set of *constructions* (short sequences of words, parts-of-speech, or constituents) from a "neutral" corpus, and then identifies constructions with distinctive distributions over the geographical subcorpora of the International Corpus of English (ICE). In social media, features of African American Vernacular English (AAVE) can be identified by correlating linguistic frequencies with the aggregate demographic statistics of the geographical areas from which geotagged social media was posted (Eisenstein et al., 2011; Stewart, 2014; Blodgett et al., 2016). In contrast, we are interested in detecting predefined dialect features from well-validated resources such as dialect atlases.

Along these lines, Jørgensen et al. (2015) and Jones (2015) designed lexical patterns to identify non-standard spellings that match known phonological variables from AAVE (e.g., *sholl* 'sure'), demonstrating the presence of these variables in social media posts from regions with high propor-

tions of African Americans. Blodgett et al. (2016) use the same geography-based approach to test for phonological spellings and constructions corresponding to syntactic variables such as habitual *be*; Hovy et al. (2015) show that a syntactic feature of Jutland Danish can be linked to the geographical origin of product reviews. These approaches have focused mainly on features that could be recognized directly from surface forms, or in some cases, from part-of-speech (POS) sequences. In contrast, we show that it is possible to learn to recognize features from examples, enabling the recognition of features for which it is difficult or impossible to craft surface or POS patterns.

Minimal pairs in NLP. A distinguishing aspect of our approach is the use of minimal pairs rather than conventional labeled data. Minimal pairs are well known in natural language processing from the Winograd Schema (Levesque et al., 2012), which is traditionally used for evaluation, but Kocijan et al. (2019) show that fine-tuning on a related dataset of minimal pairs can improve performance on the Winograd Schema itself. A similar idea arises in counterfactually-augmented data (Kaushik et al., 2019) and contrast sets (Gardner et al., 2020), in which annotators are asked to identify the minimal change to an example that is sufficient to alter its label. However, those approaches use counterfactual examples to *augment* an existing training set, while we propose minimal pairs as a replacement for large-scale labeled data. Minimal pairs have also been used to design controlled experiments and probe neural models' ability to capture various linguistic phenomena (Gulordava et al., 2018; Ettinger et al., 2018; Futrell et al., 2019; Gardner et al., 2020; Schuster et al., 2020). Finally, Liang et al. (2020) use contrastive explanations as part of an active learning framework to improve data efficiency. Our work shares the objective of Liang et al. (2020) to improve data efficiency, but is methodologically closer to probing work that uses minimal pairs to represent specific linguistic features.

7 Conclusion

We introduce the task of dialect feature detection and demonstrate that it is possible to construct dialect feature recognizers using only a small number of minimal pairs — in most cases, just five positive and negative examples per feature. This makes it possible to apply computational analysis to the many dialects for which labeled data does not exist. Future work will extend this approach to multiple dialects, focusing on cases in which features are shared across two or more dialects. This lays the groundwork for the creation of dialect-based "checklists" (Ribeiro et al., 2020) to assess the performance of NLP systems across the diverse range of linguistic phenomena that may occur in any given language.

8 Ethical Considerations

Our objective in building dialect feature recognizers is to aid developers and researchers to effectively benchmark NLP model performance across and within different dialects, and to assist social scientists and dialectologists studying dialect use. The capability to detect dialectal features may enable developers to test for and mitigate any unintentional and undesirable biases in their models towards or against individuals speaking particular dialects. This is especially important because dialect density has been documented to correlate with lower socioeconomic status (Sahgal and Agnihotri, 1988). However, this technology is not without its risks. As some dialects correlate with ethnicities or countries of origin, there is a potential dual use risk of the technology being used to profile individuals. Dialect features could also be used as predictors in downstream tasks; as with other proxies of demographic information, this could give the appearance of improving accuracy while introducing spurious correlations and imposing disparate impacts on disadvantaged groups. Hence we recommend that developers of this technology consider downstream use cases, including malicious use and misuse, when assessing the social impact of deploying and sharing this technology.

The focus on predefined dialect features can introduce a potential source of bias if the feature set is oriented towards the speech of specific subcommunities within a dialect. However, analogous issues can arise in fully data-driven approaches, in which training corpora may also be biased towards subcommunities of speakers or writers. The feature-based approach has the advantage of making any such bias easier to identify and correct.

Acknowledgments. Thanks to Claudia Lange for sharing her annotations, and for discussion of this research. Thanks to Axel Bohmann for sharing information about his work on recognizing dialect features with regular expressions. Valuable feedback on this research was provided by Jason

Baldridge, Dan Jurafsky, Slav Petrov, Jason Riesa, Kristina Toutanova, and especially Vera Axelrod. Thanks also to the anonymous reviewers. Devyani Sharma is supported in part by a Google Faculty Research Award.

References

Sarah Bunin Benor. 2010. Ethnolinguistic repertoire: Shifting the analytic focus in language and ethnicity. *Journal of Sociolinguistics*, 14(2):159–183.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving cuneiform language identification with BERT. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 17–25, Ann Arbor, Michigan. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130.

Axel Bohmann. 2019. *Variation in English worldwide: Registers and global varieties*. Cambridge University Press, Cambridge.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Holly K Craig and Julie A Washington. 2002. Oral Language Expectations for African American Preschoolers and Kindergartners. *American Journal of Speech-Language Pathology*, 11(1):59–70.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. 2000. Santa Barbara Corpus of Spoken American English. *CD-ROM. Philadelphia: Linguistic Data Consortium*.

Jonathan Dunn. 2018. Finding variants for construction-based dialectometry: A corpus-based approach to regional cxgs. *Cognitive Linguistics*, 29(2):275–311.

Jonathan Dunn. 2019. Modeling global syntactic variation in English using dialect classification. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 42–53, Ann Arbor, Michigan. Association for Computational Linguistics.

Jacob Eisenstein, Noah A. Smith, and Eric P. Xing. 2011. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1365–1374, Portland, Oregon, USA. Association for Computational Linguistics.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing Composition in Sentence Vector Representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets.

Sidney Greenbaum and Gerald Nelson. 1996. The international corpus of English (ICE) project. *World Englishes*, 15(1):3–15.

Jeffrey Grogger, Andreas Steinmayr, and Joachim Winter. 2020. The wage penalty of regional accents. NBER Working Papers 26719, National Bureau of Economic Research, Inc.

Kristina Gulordava, Piotr Bojanowski, Édouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In *Proceedings of the 2018 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1195–1205.

Dirk Hovy, Anders Johannsen, and Anders Søgaard. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, pages 452–461.

Taylor Jones. 2015. Toward a description of african american vernacular english dialect regions using "black twitter". *American Speech*, 90(4):403–440.

Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the workshop on noisy user-generated text*, pages 9–18.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2019. Learning The Difference That Makes A Difference With Counterfactually-Augmented Data. In *International Conference on Learning Representations*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A surprisingly robust trick for the winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

William Labov. 1972. *Sociolinguistic patterns*. 4. University of Pennsylvania Press.

Claudia Lange. 2012. *The syntax of spoken Indian English*. John Benjamins Publishing Company, Amsterdam.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Weixin Liang, James Zou, and Zhou Yu. 2020. Alice: Active learning with contrastive natural language explanations. *arXiv preprint arXiv:2009.10259*.

Lajanugen Logeswaran, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. Zero-shot entity linking by reading entity descriptions. *arXiv preprint arXiv:1906.07348*.

Neil A Macmillan and C Douglas Creelman. 1991. *Detection theory: A user's guide*. Cambridge University Press.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

John Nerbonne, Rinke Colen, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap-a web application for dialectology. *Dialectologia: revista electrònica*, pages 65–89.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

John R. Rickford, Greg J. Duncan, Lisa A. Gennetian, Ray Yun Gou, Rebecca Greene, Lawrence F. Katz, Ronald C. Kessler, Jeffrey R. Kling, Lisa Sanbonmatsu, Andres E. Sanchez-Ordoñez, Matthew Sciandra, Ewart Thomas, and Jens Ludwig. 2015. Neighborhood effects on use of african-american vernacular english. *Proceedings of the National Academy of Sciences*, 112(38):11817–11822.

Anju Sahgal and R. K. Agnihotri. 1988. Indian English phonology: A sociolinguistic perspective. *English World-Wide*, 9(1):51–64.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the Linguistic Signal to Predict Scalar Inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403.

Devyani Sharma. 2009. Typological diversity in New Englishes. *English World-Wide*, 30(2):170–195.

Ian Stewart. 2014. Now we stronger than ever: African-American English syntax in Twitter. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 31–37, Gothenburg, Sweden. Association for Computational Linguistics.

Janneke Van Hofwegen and Walt Wolfram. 2010. Coming of age in African American English: A longitudinal study. *Journal of Sociolinguistics*, 14(4):427–455.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings*

of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial), pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Feature	Regular expression
FOCUS *itself*	`\bitself\b`
FOCUS *only*	`\bonly\b`
NON-INITIAL EXISTENTIAL	`\bis there\b|\bare there\b`
INVARIANT TAG *(isn't it, no, na)*	`\bisn't it\b|\bis it\b|\bno\b|\bna\b`
GENERAL EXTENDER *and all*	`\band all\b`

Table 7: Regular expressions we used, for the features that such patterns were available.

A Regular Expressions

Table 7 shows the regular expressions that we used for the five features, where such patterns were available.

B Sample Outputs

The examples below represent a random sample of the multihead models' outputs for Lange's features, comparing the one that is trained on corpus examples (CORPUS) to the one that is trained on minimal pairs (MINPAIR). We show true positives (TP), false positives (FP) and false negatives (FN). We randomly sample three examples for each output type (TP, FP, FN) and model (BOTH, CORPUS only, MINPAIR only).

Our manual inspection shows a few errors in the human annotation by Lange and that certain false positives should be true positives, especially for FOCUS *only*. We highlight such examples in green. Among the rest of the false positives and false negatives, a large proportion of errors can be explained by contextual information that is not available to the models. For example, without context it is ambiguous whether "we possess only" is an example of FOCUS *only*. Inspection of context shows that it is a truncated utterance, representing a standard use of *only*, hence it is correctly characterized as a false positive. Another source of confusion to the model is missing punctuation. For example "Both girls I have never left them alone till now" could be construed as OBJECT FRONTING with RESUMPTIVE OBJECT PRONOUN. However, in the original context, the example consists of multiple sentences: "Two kids. Both girls. I have never left them alone till now." We removed punctuation from examples, since in many cases automatic ASR models do not produce punctuation either. However, this example demonstrates that punctuation can provide valuable information about clause and phrase boundaries, and should be included if possible.

B.1 Focus *itself*

[TP:BOTH] We are feeling tired now itself

[TP:BOTH] Coach means they should be coached from when they are in nursery UKG itself

[TP:BOTH] I'm in final year but like they have started from first year itself

[TP:CORPUS] And she got a chance of operating also during her internship itself nice and because that Cama hospital is for ladies only so she has lot of experience

[TP:MINPAIR] But even if they women is are working as much as a man she is earning the same monthly saving as a man itself

[TP:MINPAIR] You go around say one O'clock and then go for a movie and come back in the evening itself you see you

[FP:MINPAIR] And primarily you know the the issue orders were issued on fifth that is on the election day itself

[FP:MINPAIR] That is to we take on the coughs our human blood itself

[FP:MINPAIR] Now since you are doing the PGCT now after going back is it possible for you to use simple English in the classroom itself

[FN:BOTH] All the sums were there in the text book itself but still they have not done properly in the exam

[FN:BOTH] And thinking about dissection hall itself they really get scared and that also in the midnight

[FN:BOTH] Means what do you think that the basic itself is not good or now they are getting interest in maths

[FN:CORPUS] But even if they women is are working as much as a man she is earning the same monthly saving as a man itself

[FN:CORPUS] You go around say one O'clock and then go for a movie and come back in the evening itself you see you

[FN:MINPAIR] And she got a chance of operating also during her internship itself nice and because that Cama hospital is for ladies only so she has lot of experience

B.2 Focus *only*

[TP:BOTH] All the types only

[TP:BOTH] Hey you sur be like that only

[TP:BOTH] suddenly it will be become perfect only

[TP:CORPUS] That is I like dressing up I told you at the beginning only

[TP:CORPUS] Because today only he had come and I've got up today at nine thirty

[TP:CORPUS] Actually from childhood only I was brought up in the same atmosphere like if Papa still has shifted to another place I would have got the feeling of not having comfortable in a particular language but on the whole I think it doesn't matter exactly how we go about chosing or selecting a language

[TP:MINPAIR] it was bit it was difficult only

[TP:MINPAIR] I'm one minute I've got it in front of me only

[TP:MINPAIR] He is in our college only

[FP:BOTH] Because we are supposed to perform well there only then

[FP:BOTH] Ho Ho Hollywood Hollywood after Hollywood it seems India only

[FP:BOTH] No he'll be there in the campus only

[FP:CORPUS] Oh God there only it's happening so and forget about

[FP:CORPUS] The thing is that it is rural area only but the people are from all over india they are staying here

[FP:CORPUS] Not much work these days because first week and last week only we've quiet good business

[FP:MINPAIR] Only in India there is manual work

[FP:MINPAIR] Film hits only

[FP:MINPAIR] So Bharati Vidya Bhavan people have such type of persons only

[FN:BOTH] If they be in always that this is there are not improve no improvement only

[FN:BOTH] When we were living when I was living in Kashmir no I was brought up there only and everything is

[FN:BOTH] This is the first phase then in the second phase we have some clinical subjects in which we come in direct contact with the patients but it's on two basis like when we see the patients at the same time we study about the pathology only the pathology and then we learn about some of the drugs which are to be which are used for their treatment

[FN:CORPUS] No you must put apply science only

[FN:CORPUS] Actually they are good only

[FN:CORPUS] it was bit it was difficult only

[FN:MINPAIR] My both the parents are farmers only

[FN:MINPAIR] Because today only he had come and I've got up today at nine thirty

[FN:MINPAIR] That is I like dressing up I told you at the beginning only

B.3 Invariant Tag (*isn't it*, *no*, *na*)

[TP:BOTH] Very difficult once the school starts na very difficult

[TP:BOTH] I am okay rainy season no

[TP:BOTH] Oh yours your head is not reeling any more no ?

[TP:CORPUS] Kind of but it would be better than an indoor game no

[TP:CORPUS] We'll ask that person no that Sagar you can tell

[TP:CORPUS] Nothing at all that's why you got scratching on that day I know that no that's why I asked

[TP:MINPAIR] I'm not fair no

[TP:MINPAIR] Husband no I'll do I'll prepare it

[TP:MINPAIR] He could have agreed no what is that

[FP:BOTH] TELCO deta hai to kuch problem nahi na

[FP:BOTH] I think once you have got in you no

[FP:BOTH] I didn't go anywhere no

[FP:CORPUS] Or two hundred rupees that no

[FP:CORPUS] Know when we go back no I think we'll get a rosy welcome home welcome there

[FP:CORPUS] I like straight and perspiration then only I feel at home otherwise no

[FP:MINPAIR] No got it repaired

[FP:MINPAIR] No no he is here

[FP:MINPAIR] Okay no but

[FN:BOTH] I just go out for tea isn't

[FN:BOTH] Hey you you like serious movies is it you like serious movies

[FN:BOTH] See no the scene exactly happened you know the other day what happen I was reading baba

[FN:CORPUS] I'm not fair no

[FN:CORPUS] I think no

[FN:CORPUS] Tell me no why you can't tell

[FN:MINPAIR] Yeah then it's first time first time it was new to me no

[FN:MINPAIR] That is the main thing na here that would again the main thing that they don't take at all interest in the their children at all

[FN:MINPAIR] So culture nahi hai there is I don't follow culture religion nothing na

B.4 Lack of Copula

[FP:CORPUS] Which October first I think

[FP:CORPUS] June nineteen eighty-six

[FP:MINPAIR] Construction all before

[FP:MINPAIR] Not in the class

[FP:MINPAIR] The tendency to

[FN:BOTH] you've she said his grandfather still working

[FN:BOTH] Everybody so worried about the exams and studies

[FN:BOTH] Again classes bit too long I feel five O'clock is tiring

B.5 Left Dislocation

[TP:BOTH] This principal she is very particular about it

[TP:BOTH] Vilas and Ramesh they they make noise man

[TP:BOTH] That's why those Muslims they got very angry

[TP:CORPUS] And med medium class they can't understand soon

[TP:CORPUS] That will become difficult and common people they don't understand

[TP:CORPUS] And now the Kukis they refused to pay any more

[TP:MINPAIR] It's because of this some other participant they complained about this and then they started they started this particular

[TP:MINPAIR] We've lot of fun in theatres you know we always take the back seat and all that for this guys distinct one we keep teasing them

[TP:MINPAIR] My post graduation degree I finished it in mid June nineteen eighty-six

[FP:BOTH] But whereas when they really come to know the people they like to help the people

[FP:BOTH] It's actually some of them like to see it really so huge and long and bigger snakes they are in all closed and all there it is nice to see it

[FP:BOTH] But generally the educated people I don't find much variation but in accent there may be a variation

[FP:CORPUS] Everytime he keeps speaking you know they get irritated and say aram se

[FP:CORPUS] What happened is they will change programme and the fifty guys they'll just keep quite

[FP:CORPUS] Whereas Hyderabad the people are more conservative and like they don't like to go out even or at the first move they don't like to talk with people also

[FP:MINPAIR] And the songs now once we hear it afterwards when some other famous songs comes that we forget the last ones

[FP:MINPAIR] But when we approach since it seems they they put lot of conditions yes that you fed up with those people and

[FP:MINPAIR] so that's why we missed we that missed that holiday it being a Sunday

[FN:BOTH] Administration it is all done by Bharati Vidya Bhavan

[FN:BOTH] Oh our Joshi okay II got got him

[FN:BOTH] Yes yes it is true but our constitution makers

[FN:CORPUS] and he has used the the place where the palace once palace might be there and that portion and the remaining part he built an antenna he has fixed it there at the top

[FN:CORPUS] Not exactly but Calcutta sweets I think they do have a little flavour and that I haven't got anywhere in India

[FN:CORPUS] Computer it was in the first semester

[FN:MINPAIR] And med medium class they can't understand soon

[FN:MINPAIR] Shireen she was excellent at that

[FN:MINPAIR] Yeah arti arti students they loiter about in the corridor

B.6 Non-initial Existential *X is / are there*

[TP:BOTH] Libraries are there

[TP:BOTH] only specimen like operated cases like supposing a is there

[TP:BOTH] Problems are there problems are there what

[TP:CORPUS] to assist there some teachers are there and together we conduct the classes

[TP:CORPUS] It's there but it's common no

[TP:CORPUS] Yeah I think Varlaxmi is there

[FP:BOTH] My husband is there mother is there

[FP:CORPUS] Come no Shaukat is here Natalie is here even if Savita is not there they two are there na

[FP:CORPUS] Actually there the thing is that you know for example

[FP:CORPUS] Any thing is there produced materials which do not require much resource personnel

[FP:MINPAIR] Ph D degree is awarded there

[FN:BOTH] Yeah the royalties too there they're there and we've the king

[FN:BOTH] Okay somebody else's some somebody else is there

[FN:BOTH] In that you know everything is about nature I'll tell you yeah it's very lovely means very nice lovely what but and small children were there in that

[FN:MINPAIR] American and all other capitalist nations were also there

[FN:MINPAIR] Nice movie yaar that song is there no hai apna dil to awara

[FN:MINPAIR] It's not there

B.7 Object Fronting

[TP:BOTH] Just typing work I have to do

[TP:CORPUS] writing skills there are so many you can teach them

[TP:CORPUS] Each other and so many things we have learnt

[TP:CORPUS] My birthday party you arrange

[FP:CORPUS] Formalities I will come

[FP:CORPUS] Mar Marxism you were

[FP:MINPAIR] Other wise we have to

[FN:BOTH] That also I'm not having just I jump jumped jumped I came studies also

[FN:BOTH] Yes Hawa Mahal we heard

[FN:BOTH] About ten to twenty books I'll read that's all

[FN:MINPAIR] Small baby very nice it was

[FN:MINPAIR] But more keen she is

[FN:MINPAIR] And camera handling actually outdoor landscaping that landscape shot I have taken and actually the close ups and some parts of your architectural shots of that building Ganesh took my husband took and close ups of the faces my husband and Ganesh took

B.8 Resumptive Object Pronoun

[TP:MINPAIR] and he has used the the place where the palace once palace might be there and that portion and the remaining part he built an antenna he has fixed it there at the top

[TP:MINPAIR] Yeah also pickles we eat it with this jaggery and lot of butter

[TP:MINPAIR] My post graduation degree I finished it in mid June nineteen eighty-six

[FP:MINPAIR] Having humurous something special I would love it to join it

[FP:MINPAIR] I see a number of people I like them very much

[FP:MINPAIR] Old and ancient things in carving we get it so beautifully

[FN:BOTH] Oh our Joshi okay II got got him

[FN:BOTH] Normaly no we don't overdrawn on account but haan haan whatever is balance you know yeah help them give them suppose cheque books and all we are supposed to keep them yeah two fifty balance

[FN:BOTH] He is in a that's what he was telling me today see I want your draft like draft draft by January by the month of January by the end of January so that II might rectify it and then I will do it I will give it back to you by mid Febraury so that you can get it final draft by by the end of Febraury

[FN:CORPUS] and he has used the the place where the palace once palace might be there and that portion and the remaining part he built an antenna he has fixed it there at the top

[FN:CORPUS] Yeah also pickles we eat it with this jaggery and lot of butter

[FN:CORPUS] My post graduation degree I finished it in mid June nineteen eighty-six

B.9 Resumptive Subject Pronoun

[TP:CORPUS] Like those terrorists they wanted us to to accompany them in the revolt against India

[TP:CORPUS] And one more thing another thing how I rectified myself because all almost all all of us all my brother and sisters we have read in English medium school

[TP:CORPUS] Dr this Mr V he was totally changed actually because he was the concepts are clear not clear to us

[FP:CORPUS] There are so many people they can they could shine like anything

[FP:CORPUS] Kolhapur he had come to Guwahati

[FP:CORPUS] I don't know what he whenever whenever I see those guys they they nicely speak to me

[FP:MINPAIR] His house he is going to college KK diploma electronics

[FN:BOTH] they I thought that another one Patil is there a horrible he is I thought that Patil

[FN:BOTH] Computer it it plays a great role because we are having computers in each field now-a-days

[FN:BOTH] You know that a woman she is a apprehensive about many things

[FN:MINPAIR] Like those terrorists they wanted us to to accompany them in the revolt against India

[FN:MINPAIR] Whereas in Hyderabad they still have the old cultures and so many things that even the parents they don't even let the girls talk with the guys

[FN:MINPAIR] And the students who come out with a degree MMSI understand that there is a report that has been received from different firms that the students of BITS Pilani specially MMS candidates they are prepared to soil their hands

B.10 Topicalized Non-argument Constituent

[TP:CORPUS] for Diwali you went I know that

[TP:CORPUS] So very long time we have not travelled together

[TP:CORPUS] Pooja vacation also we used to conduct some classes practical classes

[TP:MINPAIR] In pooja day some important days we stay back

[FP:CORPUS] In Jaipur then we have also we have a Birla

[FP:CORPUS] Like that we

[FP:CORPUS] Everytime we have some work to do

[FP:MINPAIR] Aa i i initial periods I did very difficult but I

[FN:BOTH] I mean here in Hyderabad the people are it's okay they are nice

[FN:BOTH] And that old ones again we put them we feel like hearing again

[FN:BOTH] But in drama we'll have to be very different

[FN:CORPUS] In pooja day some important days we stay back

[FN:MINPAIR] for Diwali you went I know that

[FN:MINPAIR] Pooja vacation also we used to conduct some classes practical classes

[FN:MINPAIR] Sir from Monday onwards I too want to take leave sir for four days because total I have five C Ls so from

C Average Precision Results

Supervision:	**Corpus examples**		**Minimal pairs**	
Dialect feature	DAMTL	Multihead	DAMTL	Multihead
FOCUS *itself**	0.668	0.631	0.665	0.613
FOCUS *only**	0.582	0.404	0.344	0.416
INVARIANT TAG	0.876	0.871	0.441	0.495
COPULA OMISSION	0.029	0.015	0.012	0.036
LEFT DISLOCATION	0.425	0.383	0.149	0.232
NON-INITIAL EXISTENTIAL*	0.887	0.906	0.556	0.510
OBJECT FRONTING	0.238	0.202	0.031	0.083
RES. OBJECT PRONOUN	0.052	0.020	0.046	0.061
RES. SUBJECT PRONOUN	0.460	0.409	0.078	0.198
TOPICALIZED NON-ARG. CONST.	0.080	0.076	0.021	0.044
Macro Average	0.430	0.392	0.234	0.269

Table 8: Average precision for the Lange features. Scores are in the range $[0, 1]$, with 1 indicating perfect performance. Asterisks mark features that can be recognized with a regular expression.

Dialect feature	**DAMTL**	**Multihead**
ARTICLE OMISSION	0.210	0.308
DIRECT OBJECT PRO-DROP	0.044	0.057
EXTRANEOUS ARTICLE	0.116	0.065
FOCUS *itself**	1.000	0.853
FOCUS *only**	0.859	0.274
HABITUAL PROGRESSIVE	0.008	0.020
INVARIANT TAG	0.614	0.420
INVERSION IN EMBEDDED CLAUSE	0.106	0.162
LACK OF AGREEMENT	0.084	0.110
LACK OF INVERSION IN WH-QUESTIONS	0.309	0.106
LEFT DISLOCATION	0.288	0.301
MASS NOUNS AS COUNT NOUNS	0.045	0.034
NON-INITIAL EXISTENTIAL*	0.506	0.397
OBJECT FRONTING	0.147	0.193
PREPOSITION OMISSION	0.064	0.116
PP FRONTING WITH REDUCTION	0.091	0.134
STATIVE PROGRESSIVE	0.267	0.329
GENERAL EXTENDER *and all*	0.769	0.778
Macro Average	0.307	0.259

Table 9: Average precision for the extended feature set. As described in the main text, corpus training examples are unavailable for these features.

D Minimal pairs

ID	Feature	Example	Label
1	ARTICLE OMISSION	the person I like the most is from mechanical department	1
1	ARTICLE OMISSION	person I like the most is from the mechanical department	1
1	ARTICLE OMISSION	person I like most is from the mechanical department	1
1	ARTICLE OMISSION	person I like most is from mechanical department	1
1	ARTICLE OMISSION	the person I like the most is from the mechanical department	0
2	ARTICLE OMISSION	we can only see blue sky	1
2	ARTICLE OMISSION	we can only see the blue sky	0
3	ARTICLE OMISSION	recipe is simple thing	1
3	ARTICLE OMISSION	recipe is a simple thing	1
3	ARTICLE OMISSION	a recipe is simple thing	1
3	ARTICLE OMISSION	a recipe is a simple thing	0
4	ARTICLE OMISSION	union person contacted his representative at the school	1
4	ARTICLE OMISSION	the union person contacted his representative at the school	0
5	ARTICLE OMISSION	it was first day of term	1
5	ARTICLE OMISSION	it was the first day of term	0
6	DIRECT OBJECT PRO-DROP	we have two tailors who can make for us	1
6	DIRECT OBJECT PRO-DROP	we have two tailors who can make clothes for us	0
6	DIRECT OBJECT PRO-DROP	we have two tailors who can make them for us	0
7	DIRECT OBJECT PRO-DROP	he didn't give me	1
7	DIRECT OBJECT PRO-DROP	he didn't give it to me	0
8	DIRECT OBJECT PRO-DROP	in our old age we can go and enjoy	1
8	DIRECT OBJECT PRO-DROP	in our old age we can go and enjoy it	0
9	DIRECT OBJECT PRO-DROP	she doesn't like	1
9	DIRECT OBJECT PRO-DROP	she doesn't like it	0
10	DIRECT OBJECT PRO-DROP	he likes here more	1
10	DIRECT OBJECT PRO-DROP	he likes it here more	0
11	FOCUS *itself*	So if you're not good at communication you may get filtered at the first level itself	1
11	FOCUS *itself*	So if you're not good at communication you may get filtered at even the first level	0
12	FOCUS *itself*	But I did have some difficulty getting to know people among Indians itself	1
12	FOCUS *itself*	But I did have some difficulty getting to know people among Indians themselves	0
13	FOCUS *itself*	I think you should start going to the gym from now itself.	1
13	FOCUS *itself*	I think you should start going to the gym from now.	0
14	FOCUS *itself*	I did one refresher course in the month of June itself.	1
14	FOCUS *itself*	I did one refresher course in the month of June.	0
15	FOCUS *itself*	He is doing Engineering in Delhi itself.	1
15	FOCUS *itself*	He is doing Engineering in Delhi.	0
16	FOCUS *only*	I'm working very nearby to my house only	1
16	FOCUS *only*	I'm working very near my house	0
17	FOCUS *only*	recently only in April there was a big fight	1
17	FOCUS *only*	as recently as April there was a big fight	0
18	FOCUS *only*	I was there yesterday only	1
18	FOCUS *only*	I was there just yesterday	0
19	FOCUS *only*	She was brought up there and her college was there only	1
19	FOCUS *only*	She was brought up there and her college was there too	0
20	FOCUS *only*	You get on the train and buy the ticket there only	1
20	FOCUS *only*	You get on the train and buy the ticket there too	0
21	HABITUAL PROGRESSIVE	anybody giving donation, we are giving receipt	1
21	HABITUAL PROGRESSIVE	if anybody gives a donation, we give a receipt	0
22	HABITUAL PROGRESSIVE	she is getting nightmares	1
22	HABITUAL PROGRESSIVE	she gets nightmares	0
23	HABITUAL PROGRESSIVE	they are getting H1B visas to come to the country	1
23	HABITUAL PROGRESSIVE	they get H1B visas to come to the country	0
24	HABITUAL PROGRESSIVE	they are teasing the new children when they join	1
24	HABITUAL PROGRESSIVE	they tease the new children when they join	0
25	HABITUAL PROGRESSIVE	everyone is getting that vaccination in childhood	1
25	HABITUAL PROGRESSIVE	everyone gets that vaccination in childhood	0
26	INVARIANT TAG *(isn't it, no, na)*	the children are playing outside, isn't it?	1
26	INVARIANT TAG *(isn't it, no, na)*	the children are playing outside, no?	1
26	INVARIANT TAG *(isn't it, no, na)*	the children are playing outside, na?	1
26	INVARIANT TAG *(isn't it, no, na)*	the children are playing outside, aren't they?	0
27	INVARIANT TAG *(isn't it, no, na)*	I was very scared to, no?	1
27	INVARIANT TAG *(isn't it, no, na)*	I was very scared to, na?	1

27	INVARIANT TAG *(isn't it, no, na)*	I was very scared to, wasn't I?	0
28	INVARIANT TAG *(isn't it, no, na)*	the store is around the corner, no, by the post office	1
28	INVARIANT TAG *(isn't it, no, na)*	the store is around the corner, na, by the post office	1
28	INVARIANT TAG *(isn't it, no, na)*	the store is around the corner by the post office	0
29	INVARIANT TAG *(isn't it, no, na)*	It's come from me, no?	1
29	INVARIANT TAG *(isn't it, no, na)*	It's come from me, na?	1
29	INVARIANT TAG *(isn't it, no, na)*	It's come from me, hasn't it?	0
30	INVARIANT TAG *(isn't it, no, na)*	he liked it, no, even though you said he wouldn't	1
30	INVARIANT TAG *(isn't it, no, na)*	he liked it, na, even though you said he wouldn't	1
30	INVARIANT TAG *(isn't it, no, na)*	he liked it, right, even though you said he wouldn't	0
30	INVARIANT TAG *(isn't it, no, na)*	he liked it, didn't he, even though you said he wouldn't	0
31	INVERSION IN EMBEDDED CLAUSE	you cannot ask them why are they not coming for clinic visits	1
31	INVERSION IN EMBEDDED CLAUSE	you cannot ask them why they are not coming for clinic visits	0
32	INVERSION IN EMBEDDED CLAUSE	I don't know now what are they doing	1
32	INVERSION IN EMBEDDED CLAUSE	I don't know now what they are doing	0
33	INVERSION IN EMBEDDED CLAUSE	he was wondering why did the police stop him	1
33	INVERSION IN EMBEDDED CLAUSE	he was wondering why the police stopped him	0
34	INVERSION IN EMBEDDED CLAUSE	we want to know how can we make your favorite dish	1
34	INVERSION IN EMBEDDED CLAUSE	we want to know how we can make your favorite dish	0
35	INVERSION IN EMBEDDED CLAUSE	the school principal called me to ask when are you going back	1
35	INVERSION IN EMBEDDED CLAUSE	the school principal called me to ask when you are going back	0
36	LACK OF AGREEMENT	he do a lot of things	1
36	LACK OF AGREEMENT	he does a lot of things	0
37	LACK OF AGREEMENT	my bother said that one of his favorite place is the beach nearby	1
37	LACK OF AGREEMENT	my bother said that one of his favorite places is the beach nearby	0
38	LACK OF AGREEMENT	only his shoes is visible	1
38	LACK OF AGREEMENT	only his shoes are visible	0
39	LACK OF AGREEMENT	ten years ago you didn't operated a machine that could lift a house all by itself	1
39	LACK OF AGREEMENT	ten years ago you didn't operate a machine that could lift a house all by itself	0
40	LACK OF AGREEMENT	he talk to them	1
40	LACK OF AGREEMENT	he talks to them	0
41	LACK OF INV. IN WH-QUESTIONS	where you will get anything?	1
41	LACK OF INV. IN WH-QUESTIONS	where will you get anything?	0
42	LACK OF INV. IN WH-QUESTIONS	what you are doing?	1
42	LACK OF INV. IN WH-QUESTIONS	what are you doing?	0
43	LACK OF INV. IN WH-QUESTIONS	why you are telling this to everybody?	1
43	LACK OF INV. IN WH-QUESTIONS	why are you telling this to everybody?	0
44	LACK OF INV. IN WH-QUESTIONS	why you are driving like a lorry?	1
44	LACK OF INV. IN WH-QUESTIONS	why are you driving like a lorry?	0
45	LACK OF INV. IN WH-QUESTIONS	how your mother is feeling?	1
45	LACK OF INV. IN WH-QUESTIONS	how is your mother feeling?	0
46	LEFT DISLOCATION	my father, he works for a mining company	1
46	LEFT DISLOCATION	my father works for a mining company	0
47	LEFT DISLOCATION	nowadays all the children they are mature from a very early age	1
47	LEFT DISLOCATION	nowadays all the children are mature from a very early age	0
48	LEFT DISLOCATION	the camera, the dog is facing towards it	1
48	LEFT DISLOCATION	the dog is facing towards the camera	0
49	LEFT DISLOCATION	and all the company people, they are my clients	1
49	LEFT DISLOCATION	and all the company people are my clients	0
50	LEFT DISLOCATION	those who come here definitely they should learn English	1
50	LEFT DISLOCATION	those who come here should definitely learn English	0
51	MASS NOUNS AS COUNT NOUNS	this is a menial work	1
51	MASS NOUNS AS COUNT NOUNS	this is menial work	0
52	MASS NOUNS AS COUNT NOUNS	open a shop wherever there is a foot traffic	1
52	MASS NOUNS AS COUNT NOUNS	open a shop wherever there is foot traffic	0
53	MASS NOUNS AS COUNT NOUNS	all the musics are very good	1
53	MASS NOUNS AS COUNT NOUNS	all the music is very good	0
54	MASS NOUNS AS COUNT NOUNS	some informations are available free	1
54	MASS NOUNS AS COUNT NOUNS	some information is available free	0
55	MASS NOUNS AS COUNT NOUNS	they use proper grammars there	1
55	MASS NOUNS AS COUNT NOUNS	they use proper grammar there	0
56	NON-INITIAL EXISTENTIAL	some flower part is there	1
56	NON-INITIAL EXISTENTIAL	there is some flower part	0
57	NON-INITIAL EXISTENTIAL	corruption is there obviously	1
57	NON-INITIAL EXISTENTIAL	there is corruption obviously	0
58	NON-INITIAL EXISTENTIAL	because in India individuality is not there	1
58	NON-INITIAL EXISTENTIAL	because there is no individuality in India	0

59	NON-INITIAL EXISTENTIAL	five balls are there	1
59	NON-INITIAL EXISTENTIAL	there are five balls	0
60	NON-INITIAL EXISTENTIAL	every year inflation is there	1
60	NON-INITIAL EXISTENTIAL	every year there is inflation	0
61	OBJECT FRONTING	not so much adjustment i have to make	1
61	OBJECT FRONTING	i don't have to make so much adjustment	0
61	OBJECT FRONTING	i have to make not so much adjustment	0
62	OBJECT FRONTING	minimum one month you have to wait	1
62	OBJECT FRONTING	you have to wait a minimum of one month	0
63	OBJECT FRONTING	Hindi Gujarati and Marathi you can use in Bombay	1
63	OBJECT FRONTING	you can use Hindi Gujarati and Marathi in Bombay	0
64	OBJECT FRONTING	in fifteen years lot of changes we have seen	1
64	OBJECT FRONTING	in fifteen years we have seen a lot of changes	0
65	OBJECT FRONTING	tomorrow this cake you have to try	1
65	OBJECT FRONTING	tomorrow you have to try this cake	0
66	PREPOSITION OMISSION	I can see some green colour leaves the left side	1
66	PREPOSITION OMISSION	I can see some green colour leaves on the left side	0
67	PREPOSITION OMISSION	I went one year there.	1
67	PREPOSITION OMISSION	I went there for one year.	0
68	PREPOSITION OMISSION	We don't feel that we should go any other country.	1
68	PREPOSITION OMISSION	We don't feel that we should go to any other country.	0
69	PREPOSITION OMISSION	Those days it was considered a good job.	1
69	PREPOSITION OMISSION	In those days it was considered a good job.	0
70	PREPOSITION OMISSION	So that time they said okay go and work for a few months.	1
70	PREPOSITION OMISSION	So at that time they said okay go and work for a few months.	0
71	PP FRONTING WITH REDUCTION	first of all, right side we can see a plate	1
71	PP FRONTING WITH REDUCTION	first of all, we can see a plate the right side	0
71	PP FRONTING WITH REDUCTION	first of all, we can see a plate on the right side	0
71	PP FRONTING WITH REDUCTION	first of all, on the right side we can see a plate	0
71	ARTICLE OMISSION	first of all, right side we can see a plate	1
71	PREPOSITION OMISSION	first of all, right side we can see a plate	1
71	PREPOSITION OMISSION	first of all, we can see a plate the right side	1
72	PP FRONTING WITH REDUCTION	Tirupati temple I stayed one or two days	1
72	PP FRONTING WITH REDUCTION	I stayed one or two days at the Tirupati temple	0
72	PP FRONTING WITH REDUCTION	at the Tirupati temple I stayed one or two days	0
72	ARTICLE OMISSION	Tirupati temple I stayed one or two days	1
72	PREPOSITION OMISSION	Tirupati temple I stayed one or two days	1
73	PP FRONTING WITH REDUCTION	two years I stayed alone	1
73	PP FRONTING WITH REDUCTION	for two years I stayed alone	0
73	PREPOSITION OMISSION	two years I stayed alone	1
74	PP FRONTING WITH REDUCTION	you can say anything but tenth I'm leaving	1
74	PP FRONTING WITH REDUCTION	you can say anything but on the tenth I'm leaving	0
74	ARTICLE OMISSION	you can say anything but tenth I'm leaving	1
74	PREPOSITION OMISSION	you can say anything but tenth I'm leaving	1
75	PP FRONTING WITH REDUCTION	actually, this part I have not been	1
75	PP FRONTING WITH REDUCTION	actually, I have not been to this part	0
75	PREPOSITION OMISSION	actually, this part I have not been	1
76	STATIVE PROGRESSIVE	they are speaking Portuguese in Brazil	1
76	STATIVE PROGRESSIVE	they speak Portuguese in Brazil	0
77	STATIVE PROGRESSIVE	and the production function is giving you the relationship between input and output	1
77	STATIVE PROGRESSIVE	and the production function gives you the relationship between input and output	0
77	ARTICLE OMISSION	and the production function is giving you the relationship between input and output	1
77	ARTICLE OMISSION	and the production function gives you the relationship between input and output	1
78	STATIVE PROGRESSIVE	he is having a television	1
78	STATIVE PROGRESSIVE	he has a television	0
79	STATIVE PROGRESSIVE	I think Nina must be knowing her sister	1
79	STATIVE PROGRESSIVE	I think Nina must know her sister	0
80	STATIVE PROGRESSIVE	we will be knowing how much the structure is getting deflected	1
80	STATIVE PROGRESSIVE	we will know how much the structure is getting deflected	0
81	EXTRANEOUS ARTICLE	Chandigarh was full of the employed people.	1
81	EXTRANEOUS ARTICLE	Chandigarh was full of employed people.	0
82	EXTRANEOUS ARTICLE	She has a business experience.	1
82	EXTRANEOUS ARTICLE	She has business experience.	0
83	EXTRANEOUS ARTICLE	Because educated people get a good money.	1
83	EXTRANEOUS ARTICLE	Because educated people get good money.	0

84	EXTRANEOUS ARTICLE	They have a pressure from their in-laws.	1
84	EXTRANEOUS ARTICLE	They have pressure from their in-laws.	0
85	EXTRANEOUS ARTICLE	Here the life is busy.	1
85	EXTRANEOUS ARTICLE	Here life is busy.	0
86	GENERAL EXTENDER *and all*	So marketing keeps its communication with the different embassies and all.	1
86	GENERAL EXTENDER *and all*	So marketing keeps its communication with the different embassies.	0
87	GENERAL EXTENDER *and all*	Whereas we had lot of time and we didn't have any TV and all and we used to play outdoor games.	1
87	GENERAL EXTENDER *and all*	Whereas we had lot of time and we didn't have any TV and we used to play outdoor games.	0
87	GENERAL EXTENDER *and all*	Whereas we had lot of time and we didn't have any TV and all that stuff and we used to play outdoor games.	0
88	GENERAL EXTENDER *and all*	So I did my schooling and all from there.	1
88	GENERAL EXTENDER *and all*	So I did my schooling from there.	0
89	GENERAL EXTENDER *and all*	We are like we are in touch, but not before when we was in school and all.	1
89	GENERAL EXTENDER *and all*	We are like we are in touch, but not before when we was in school.	0
89	LACK OF AGREEMENT	We are like we are in touch, but not before when we was in school and all.	1
89	LACK OF AGREEMENT	We are like we are in touch, but not before when we was in school.	1
90	GENERAL EXTENDER *and all*	My parents and siblings and all, they really enjoy playing board games.	1
90	GENERAL EXTENDER *and all*	My parents and siblings, they really enjoy playing board games.	0
90	LEFT DISLOCATION	My parents and siblings and all, they really enjoy playing board games.	1
90	LEFT DISLOCATION	My parents and siblings, they really enjoy playing board games.	1
91	COPULA OMISSION	I think she a teacher.	1
91	COPULA OMISSION	I think she is a teacher.	0
92	COPULA OMISSION	They all aggressive states.	1
92	COPULA OMISSION	They are all aggressive states.	0
93	COPULA OMISSION	Now they wearing American type of dresses.	1
93	COPULA OMISSION	Now they are wearing American type of dresses.	0
94	COPULA OMISSION	So my parents from Gujarat.	1
94	COPULA OMISSION	So my parents are from Gujarat.	0
95	COPULA OMISSION	Sorry I can't come, everything busy in our life.	1
95	COPULA OMISSION	Sorry I can't come, everything is busy in our life.	0
96	RESUMPTIVE OBJECT PRONOUN	The cake, I like it very much.	1
96	RESUMPTIVE OBJECT PRONOUN	I like the cake very much.	0
96	LEFT DISLOCATION	The cake, I like it very much.	1
97	RESUMPTIVE OBJECT PRONOUN	The book that I left it here, where is it?	1
97	RESUMPTIVE OBJECT PRONOUN	The book that I left here, where is it?	0
98	RESUMPTIVE OBJECT PRONOUN	My old life I want to spend it in India.	1
98	RESUMPTIVE OBJECT PRONOUN	My old life I want to spend in India.	0
98	LEFT DISLOCATION	My old life I want to spend it in India.	1
99	RESUMPTIVE OBJECT PRONOUN	Some teachers when I was in school I liked them very much.	1
99	RESUMPTIVE OBJECT PRONOUN	Some teachers when I was in school I liked very much.	0
99	LEFT DISLOCATION	Some teachers when I was in school I liked them very much.	1
100	RESUMPTIVE OBJECT PRONOUN	I'm going to find my bag which I left it in the room.	1
100	RESUMPTIVE OBJECT PRONOUN	I'm going to find my bag which I left in the room.	0
101	RESUMPTIVE SUBJECT PRONOUN	A person living in Calcutta, which he didn't know Hindi earlier, when he comes to Delhi he has to learn English.	1
101	RESUMPTIVE SUBJECT PRONOUN	A person living in Calcutta, who didn't know Hindi earlier, when he comes to Delhi he has to learn English.	0
102	RESUMPTIVE SUBJECT PRONOUN	But now all kids they have a computer and all new technology.	1
102	RESUMPTIVE SUBJECT PRONOUN	But now all kids have a computer and all new technology.	0
102	LEFT DISLOCATION	But now all kids they have a computer and all new technology.	1
103	RESUMPTIVE SUBJECT PRONOUN	My daughter she is attending the University of Delhi.	1
103	RESUMPTIVE SUBJECT PRONOUN	My daughter is attending the University of Delhi.	0
103	LEFT DISLOCATION	My daughter she is attending the University of Delhi.	1
104	RESUMPTIVE SUBJECT PRONOUN	and that roommate, he will do an interview	1

104	RESUMPTIVE SUBJECT PRONOUN	and that roommate will do an interview	0
104	LEFT DISLOCATION	and that roommate, he will do an interview	1
105	RESUMPTIVE SUBJECT PRONOUN	some people they are very nice	1
105	RESUMPTIVE SUBJECT PRONOUN	some people are very nice	0
105	LEFT DISLOCATION	some people they are very nice	1
106	TOPICALIZED NON-ARG. CONST	daytime I work for the courier service	1
106	TOPICALIZED NON-ARG. CONST	in the daytime I work for the courier service	1
106	TOPICALIZED NON-ARG. CONST	I work for the courier service in the daytime	0
106	PP FRONTING WITH REDUCTION	daytime I work for the courier service	1
107	TOPICALIZED NON-ARG. CONST	for many years I did not travel	1
107	TOPICALIZED NON-ARG. CONST	many years I did not travel	1
107	TOPICALIZED NON-ARG. CONST	I did not travel for many years	0
107	PP FRONTING WITH REDUCTION	many years I did not travel	1
108	TOPICALIZED NON-ARG. CONST	with your mother I love to go shopping	1
108	TOPICALIZED NON-ARG. CONST	I love to go shopping with your mother	0
109	TOPICALIZED NON-ARG. CONST	and in the background there are a lot of buildings	1
109	TOPICALIZED NON-ARG. CONST	and there are a lot of buildings in the background	0
110	TOPICALIZED NON-ARG. CONST	yeah, so my parent's house I go very often	1
110	TOPICALIZED NON-ARG. CONST	yeah, so to my parent's house I go very often	1
110	TOPICALIZED NON-ARG. CONST	yeah, so I go very often to my parent's house	0
110	PP FRONTING WITH REDUCTION	yeah, so my parent's house I go very often	1