The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges

Genta Indra Winata[1], **Alham Fikri Aji**[2], **Zheng-Xin Yong**[3], **Thamar Solorio**[1]*
[1]Bloomberg [2]MBZUAI [3]Brown University
gwinata@bloomberg.net, alham.fikri@mbzuai.ac.ae, contact.yong@brown.edu

Abstract

Code-Switching, a common phenomenon in written text and conversation, has been studied over decades by the natural language processing (NLP) research community. Initially, code-switching is intensively explored by leveraging linguistic theories and, currently, more machine-learning oriented approaches to develop models. We introduce a comprehensive systematic survey on code-switching research in natural language processing to understand the progress of the past decades and conceptualize the challenges and tasks on the code-switching topic. Finally, we summarize the trends and findings and conclude with a discussion for future direction and open questions for further investigation.

1 Introduction

Code-Switching is the linguistic phenomenon where multilingual speakers use more than one language in the same conversation (Poplack, 1978). The fragment of the worldwide population that can be considered multilingual, i.e., speaks more than one language, far outnumbers monolingual speakers (Tucker, 2001; Winata et al., 2021a). This alone makes a compelling argument for developing NLP technology that can successfully process code-switched (CSW) data. However, it was not until the last couple of years that CSW-related research became more popular (Sitaram et al., 2019; Jose et al., 2020; Doğruöz et al., 2021), and this increased interest has been motivated to a large extent by: 1) The need to process social media data. Before the proliferation of social media platforms, it was more common to observe code-switching in spoken language and not so much in written language. This is not the case anymore, as multilingual users tend to combine the languages they speak on social media; 2) The increasing release of voice-operated devices. Now that smart assistants are becoming more and more accessible, we have started to realize that assuming users will interact with NLP technology as monolingual speakers is very restrictive and does not fulfill the needs of real-world users. Multilingual speakers also prefer to interact with machines in a CSW manner (Bawa et al., 2020). We show quantitative evidence of the upward trend for CSW-related research in Figure 1.

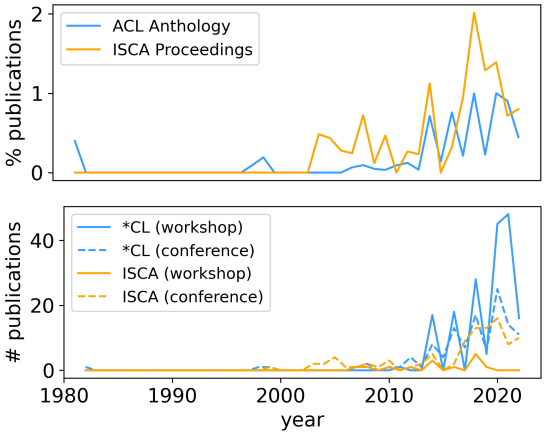


Figure 1: Number of publications over time in *CL and ISCA venues. We collect the papers on October 2022. **Top:** Relative to all *CL and ISCA papers. **Bottom:** absolute number, broken down into conferences vs workshops. It does not include papers published after. The graphs do not show the number of publications published in journals and symposiums.

In this paper, we present the first large-scale comprehensive survey on CSW NLP research in a structured manner by collecting more than 400 papers published on open repositories, such as the ACL Anthology and ISCA proceedings (see §2). We manually coded these papers to collect coarse- and fine-grained information (see §2.1) on CSW research in NLP that includes languages covered (see §3), NLP tasks that have been explored, and new and emerging trends (see §4). In addition, motivated by the fact that fields like linguistics, socio-linguistics, and related fields, have studied

* The work was done while at Bloomberg.

Findings of the Association for Computational Linguistics: ACL 2023, pages 2936–2978
July 9-14, 2023

Category	Options
Languages	Bilingual, Trilingual, 4+
Venues	Conference, Workshop, Symposium, Book
Papers	Theory / Linguistics, Empirical, Analysis, Position/Opinion/Survey, Metric, Corpus, Shared Task, Demo
Datasets	Social Media, Speech (Recording), Transcription, News, Dialogue, Books, Government Document, Treebank
Methods	Rule/Linguistic Constraint, Statistical Model, Neural Network, Pre-trained Model
Tasks	**Text:** Topic Modeling, Semantic Parsing, Dependency Parsing, Sentiment Analysis, Emotion Detection, Abusive Language Detection, Sarcasm Detection, Humor Detection, Humor Generation, Dialogue State Tracking, Text Generation, Natural Language Understanding, Named Entity Recognition, Part-of-Speech Tagging, Natural Language Entailment, Language Modeling, Regression, Language Identification, Machine Translation, Text Normalization, Micro-Dialect Identification, Question Answering, Summarization **Speech:** Acoustic Modeling, Speech Recognition, Text-to-Speech, Speech Synthesis

Table 1: Categories in the annotation scheme.

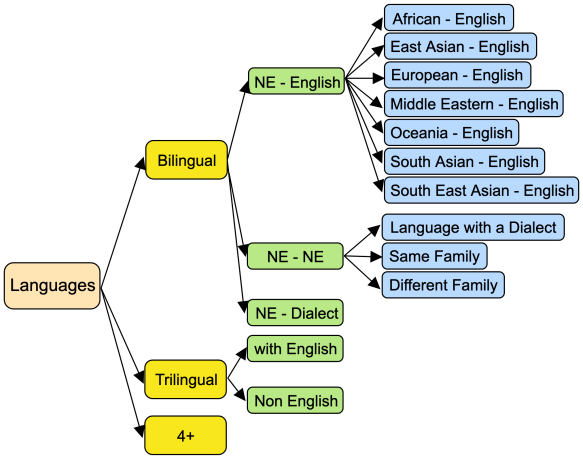


Figure 2: Language Categories. *NE denotes Non English. We show fine-grained categories in **green** and **blue**.

CSW since the early 1900s, we also investigate to what extent theoretical frameworks from these fields have influenced NLP approaches (see §5), and how the choice of methods has evolved over time (see §5.4). Finally, we discuss the most pressing research challenges and identify a path forward to continue advancing this exciting line of work (see §6).

The area of NLP for CSW data is thriving, covering an increasing number of language combinations and tasks, and it is clearly advancing from a niche field to a common research topic, thus making our comprehensive survey timely. We expect the survey to provide valuable information to researchers new to the field and motivate more research from researchers already engaging in NLP for CSW data.

2 Exploring Open Proceedings

To develop a holistic understanding of the trends and advances in CSW NLP research, we collect research papers on CSW from the ACL Anthology and ISCA proceedings. We focus on these two sources because they encompass the top venues for publishing in speech and language processing in our field. In addition, we also look into personal repositories from researchers in the community that contains a curated list of CSW-related papers. We discuss below the search process for each venue.

ACL Anthology We crawled the entire the ACL Anthology repository up to October 2022.[1] We then filtered papers by using the following keywords related to CSW: “codeswitch", “code switch", “code-switching", “code-switched", “code-switch", “code-mix", “code-mixed", “code-mixing", “code mix" “mixed-language", “mixed-lingua", “mixed language", “mixed lingua", and “mix language".

ISCA Proceedings We manually reviewed publicly available proceedings on the ISCA website[2] and searched for papers related to CSW using the same set of keywords as above.

Web Resources To extend the coverage of paper sources, we also gathered data from existing repositories.[3,4] We can find multiple linguistics papers studying about CSW.

2.1 Annotation Process

We have three human annotators to annotate all collected papers based on multiple categories shown in Table 1. All papers are coded by a least one

[1] https://github.com/acl-org/acl-anthology

[2] https://www.isca-speech.org

[3] https://github.com/gentaiscool/code-switching-papers

[4] https://genius1237.github.io/emnlp19_tut/

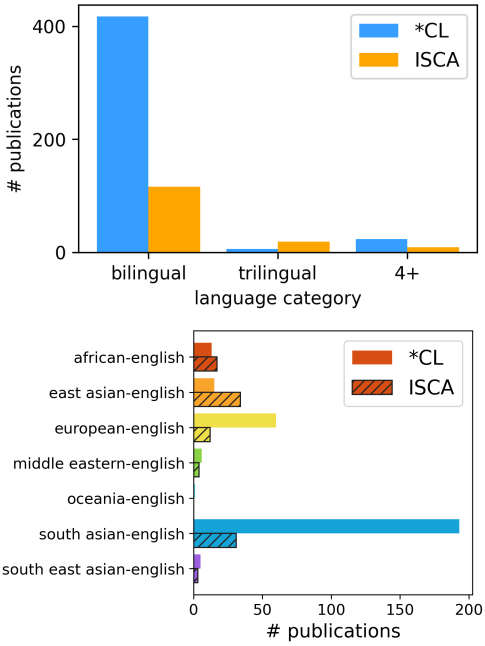


Figure 3: **(Top):** Number of publications across the type of language combination (bilingual, trilingual or 4+. **(Bottom):** Number of publications on fine-grained bilingual category with English as the L2 language.

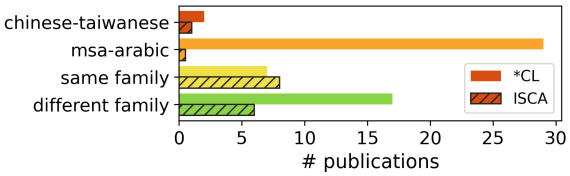


Figure 4: Number of publications of bilingual code-switched languages that do not contain English. *msa stands for Modern Standard Arabic. The first two are the combination of a language with its dialect.

annotator. To extract the specific information we are looking for from the paper, the annotator needs to read through the paper, as most of the information is not contained in the abstract. The full list of the annotations we collected is available in the Appendix (see §A).

To facilitate our analysis, we annotated the following aspects:

- **Languages:** CSW is not restricted to pairs of languages; thus, we divide papers by the number of languages that are covered into `bilingual`, `trilingual`, and `4+` (if there are at least four languages). For a more fine-grained classification of languages, we categorize them by geographical location (see Figure 2).

Languages	# Publications			
	*CL	ISCA	Total	Shared Task
Hindi-English	111	17	128	30
Spanish-English	78	8	86	40
Chinese-English‡	20	27	47	5
Tamil-English	37	2	39	17
Malayalam-English	23	2	25	13

Table 2: Most common code-switching languages in *CL and ISCA venues. ‡The count does not include the dialect or South East Asian Mandarin - English (Lyu et al., 2010a) since they contain more than two languages (i.e., it has words from the Chinese dialect).

- **Venues:** There are multiple venues for CSW-related publications. We considered the following type of venues: conference, workshop, symposium and book. As we will discuss later, the publication venue is a reasonable data point of the wider spread of CSW research in recent years.

- **Papers:** We classify the paper types based on their contribution and nature. We predict that we will have a high distribution of dataset/resource papers, as lack of resources has been a major bottleneck in the past.

- **Datasets:** If the paper uses a dataset for the research, we will identify the source and modality (i.e., written text or speech) of the dataset.

- **Methods:** We identify the type of methods presented in work.

- **Tasks:** We identify the downstream NLP tasks (including the speech processing-related tasks) presented in work.

3 Language Diversity

Here, we show the languages covered in the CSW resources. While focusing on the CSW phenomenon increases diversity of NLP technology, as we will see in this section, future efforts are needed to provide significant coverage of the most common CSW language combinations worldwide.

3.1 Variety of Mixed Languages

Figure 3 shows the distribution of languages represented in the NLP for CSW literature. Most of the papers use datasets with two language pairs. However, we did find a few papers that address CSW scenarios with more than two languages. We

Languages	# Publications		
	non-ST	ST	Total
Language Identification	46	17	63
Sentiment Analysis	31	30	61
NER	17	14	31
POS Tagging	29	1	30
Abusive/Offensive Lang. Detection	9	16	25
ASR	20	0	22
Language Modeling	19	1	20
Machine Translation	8	5	13

Table 3: Most common tasks in ACL venues. ST denotes shared task.

consider this a relevant future direction in CSW: scaling model abilities to cover n languages, with $n \geq 2$.

CSW in two languages We group the number of publications focusing on bilingual CSW based on world regions in Figure 3 (bottom). We can see that the majority of research in CSW has focused on South Asian-English, especially on Hindi-English, Tamil-English, and Malayalam-English, as shown in Table 2. The other common language pairs are Spanish-English and Chinese-English. That table also shows that many of the publications are shared task papers. This is probably reflecting efforts from a few research groups to motivate more research into CSW, such as that behind the CALCS workshop series.

Looking at the languages covered, we also find that there are many language pairs that come from different language families, such as Turkish-German (Çetinoğlu, 2016; Çetinoğlu and Çöltekin, 2019; Özateş and Çetinoğlu, 2021; Özateş et al., 2022), Turkish-Dutch (Gambäck and Das, 2016), French-Arabic (Sankoff, 1998; Lounnas et al., 2021), Russian-Tatar (Taguchi et al., 2021), Russian-Kazakh (Mussakhojayeva et al., 2022a), Hindi-Tamil (Thomas et al., 2018b), Arabic-North African (El-Haj et al., 2018), Basque-Spanish (Aguirre et al., 2022), and Wixarika-Spanish (Mager et al., 2019). There are only very few papers working on Middle Eastern - English language pairs, most of the time, the Middle Eastern languages are mixed with non-English and/or dialects of these languages (see Figure 4).

Trilingual The number of papers addressing CSW in more than two languages is still small (see 3 top), compared to the papers looking at pairs of languages. Not surprisingly, this smaller number of papers focus on world regions where either the official languages are more than two, or these languages are widely used in the region, for example, Arabic-English-French (Abdul-Mageed et al., 2020), Hindi-Bengali-English (Barman et al., 2016), Tulu-Kannada-English (Hegde et al., 2022), and Darija-English-French (Voss et al., 2014).

	# Publications		
	*CL	ISCA	Total
Public Dataset	38	4	42
Private Dataset	54	18	72

Table 4: Publications that introduce new corpus.

4+ When looking at the papers that focus on more than three languages, we found that many papers use South East Asian Mandarin-English (SEAME) dataset (Lyu et al., 2010a), which has Chinese dialects and Malay or Indonesian words. Most of the other datasets are machine-generated using rule-based or neural methods.

3.2 Language-Dialect Code-Switching

Based on Figure 4, we can find some papers with language-dialect CSW, such as Chinese-Taiwanese Dialect (Chu et al., 2007; Yu et al., 2012) and Modern Standard Arabic (MSA)-Arabic Dialect (Elfardy and Diab, 2012; Samih and Maier, 2016; El-Haj et al., 2018). The dialect, in this case, is the variation of the language with a different form that is very specific to the region where the CSW style is spoken.

4 Tasks and Datasets

In this section, we summarize our findings, focusing on the CSW tasks and datasets. Table 3 shows the distribution of CSW tasks for ACL papers with at least ten publications. The two most popular tasks are language identification and sentiment analysis. Researchers mostly use the shared tasks from 2014 (Solorio et al., 2014) and 2016 (Molina et al., 2016) for language identification, and the SemEval 2020 shared task (Patwa et al., 2020) for sentiment analysis. For ISCA, the most popular tasks are unsurprisingly ASR and TTS. This strong correlation between task and venue shows that the speech processing and *CL communities remain somehow fragmented and working in isolation from one another, from the most part.

Public vs. Private Datasets Public datasets availability also dictates what tasks are being

Source	*CL	ISCA	Total
Social Media	183	3	186
Speech (Recording)	29	102	141
Transcription	23	4	27
News	19	5	24
Dialogue	16	2	18
Books	7	1	8
Government Document	6	0	6
Treebank	5	0	5

Table 5: The source of the CSW dataset in the literature.

Type	*CL	ISCA	Total
Empirical	205	100	305
Shared Task	82	1	83
Corpus (Closed)	54	18	62
Corpus (Open)	38	4	42
Analysis	34	8	42
Demo	7	2	9
Theoretical/Linguistic	7	0	7
Position/Opinion/Survey	3	0	3
Metric	2	1	3

Table 6: Paper Type of the CSW papers.

explored in CSW research. Public datasets such as HinGE (Srivastava and Singh, 2021b), SEAME (Lyu et al., 2010a) and shared task datasets (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018; Patwa et al., 2020) have been widely used in many of the papers. Some work, however, used new datasets that are not publicly available, thus hindering adoption (see Table 4). There are two well-known benchmarks in CSW: LinCE (Aguilar et al., 2020) and GlueCOS (Khanuja et al., 2020b). These two benchmarks have a handful of tasks, and they are built to encourage transparency and reliability of evaluation since the test set labels are not publicly released. The evaluation is done automatically on their websites. However, their support languages are mostly limited to popular CSW language pairs, such as Spanish-English, Modern Standard Arabic-Egyptian, and Hindi-English, the exception being Nepali-English in LinCE.

Dataset Source Table 5 shows the statistics of dataset sources in the CSW literature. We found that most of the ACL papers were working on social media data. This is expected, considering that social media platforms are known to host informal interactions among users, making them reasonable places for users to code-switch. Naturally, most ISCA papers work on speech data, many of which are recordings of conversations and interviews. There are some datasets that come from speech transcription, news, dialogues, books, government documents, and treebanks.

Paper Category Table 6 presents the distribution of CSW papers. Most of the papers are empirical work focusing on the evaluation of downstream tasks. The second largest population is shared tasks. We also notice that many papers introduce new CSW corpus, but they are not released publicly. Some papers only release the URL or id to download the datasets, especially for datasets that come from social media (e.g., Twitter) since redistribution of the actual tweets is not allowed (Solorio et al., 2014; Molina et al., 2016) resulting in making reproducibility harder. Social media users can delete their posts at any point in time, resulting in considerable data attrition rates. There are very few papers working on the demos, theoretical work, position papers, and introducing evaluation metrics.

5 From Linguistics to NLP

Notably, papers are working on approaches that are inspired by linguistic theories to enhance the processing of CSW text. In this survey, we find three linguistic constraints that are used in the literature: equivalence constraint, matrix-embedded language Framework (MLF), and Functional Head Constraint. In this section, we will briefly introduce the constraints and list the papers that utilize the constraints.

5.1 Linguistic-Driven Approaches

Equivalence Constraint In a well-formed code-switched sentence, the switching takes place at those points where the grammatical constraints of both languages are satisfied (Poplack, 1980). Li and Fung (2012, 2013) incorporate this syntactic constraint to a statistical code-switch language model (LM) and evaluate the model on Chinese-English code-switched speech recognition. On the same line of work, Pratapa et al. (2018a); Pratapa and Choudhury (2021) implement the same constraint to Hindi-English CSW data by producing parse trees of parallel sentences and matching the surface order of child nodes in the trees. Winata et al. (2019c) apply the constraint to generate synthetic CSW text and find that combining the real

CSW data with synthetic CSW data can effectively improve the perplexity. They also treat parallel sentences as a linear structure and only allow switching on non-crossing alignments.

Matrix-Embedded Language Framework (MLF) Myers-Scotton (1997) proposed that in bilingual CSW, there exists an asymmetrical relationship between the dominant *matrix language* and the subordinate *embedded language*. Matrix language provides the frame of the sentence by governing all or most of the most of the grammatical morphemes as well as word order, whereas syntactic elements that bear no or only limited grammatical function can be provided by the embedded language (Johanson, 1999; Myers-Scotton, 2005). Lee et al. (2019a) use augmented parallel data by utilizing MLF to supplement the real code-switch data. Gupta et al. (2020) use MLF to automatically generate the code-mixed text from English to multiple languages without any parallel data.

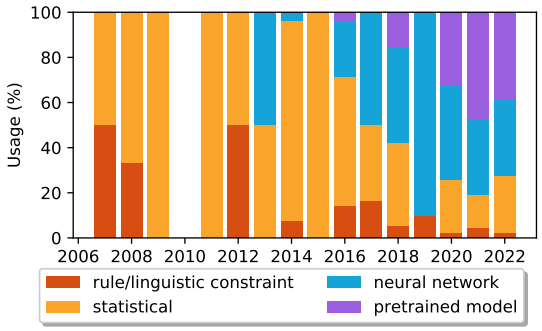
Functional Head Constraint Belazi et al. (1994) posit that it is impossible to switch languages between a functional head and its complement because of the strong relationship between the two constituents. Li and Fung (2014) use the constraint of the LM by first expanding the search network with a translation model and then using parsing to restrict paths to those permissible under the constraint.

5.2 Learning from Data Distribution

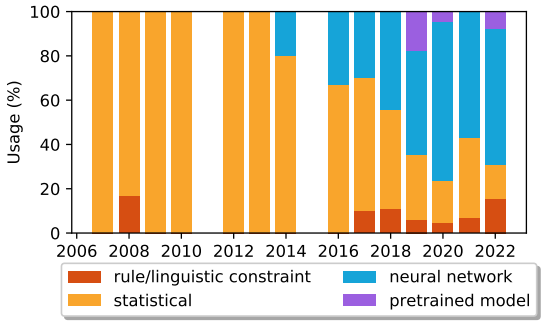
Linguistic constraint theories have been used for decades to generate synthetic CSW sentences to address the lack of data issue. However, the approach requires external word alignments or constituency parsers that create erroneous results instead of applying the linguistic constraints to generate new synthetic CSW data, building a pointer-generator model to learn the real distribution of code-switched data (Winata et al., 2019c). Chang et al. (2019) propose to generate CSW sentences from monolingual sentences using Generative Adversarial Network (GAN) (Goodfellow et al., 2020) and the generator learns to predict CSW points without any linguistic knowledge.

5.3 The Era of Statistical Methods

The research on CSW is also influenced by the progress and development of machine learning.



(a) *CL



(b) ISCA

Figure 5: Methods used for code-mixing NLP.

According to Figure 5, starting in 2006, statistical methods have been adapted to CSW research, while before that year, the approaches were mainly rule-based. There are common statistical methods for text classification used in the literature, such as Naive Bayes (Solorio and Liu, 2008a) and Support Vector Machine (SVM) (Solorio and Liu, 2008b). Conditional Random Field (CRF) (Sutton et al., 2012) is also widely seen in the literature for sequence labeling, such as Part-of-Speech (POS) tagging (Vyas et al., 2014), Named Entity Recognition (NER), and word-level language identification (Lin et al., 2014; Chittaranjan et al., 2014; Jain and Bhat, 2014). HMM-based models have been used in speech-related tasks, such as speech recognition (Weiner et al., 2012a; Li and Fung, 2013) and text synthesis (Qian et al., 2008; Shuang et al., 2010; He et al., 2012).

5.4 Utilizing Neural Networks

Following general NLP trends, we see the adoption of neural methods and pre-trained models growing in popularity over time. In contrast, the statistical and rule-based approaches are diminishing. Compared to ISCA, we see more adaptation of the pre-training model. This is because ACL work is more text-based focused, where pre-trained LMs

are more widely available.

Neural-Based Models Figure 5 shows that the trend of using neural-based models started in 2013, and the usage of rule/linguistic constraint and statistical methods diminished gradually through time, but they are still used even with a low percentage. RNN and LSTM architectures are commonly used in sequence modeling, such as language modeling (Adel et al., 2013; Vu and Schultz, 2014; Adel et al., 2014c; Winata et al., 2018a; Garg et al., 2018a; Winata et al., 2019c) and CSW identification (Samih et al., 2016a). DNN-based and hybrid HMM-DNN models are used in speech recognition models (Yilmaz et al., 2018; Yılmaz et al., 2018).

Pre-trained Embeddings Pre-trained embeddings are used to complement neural-based approaches by initializing the embedding layer. Common pre-trained embeddings used in the literature are monolingual subword-based embeddings, FastText (Joulin et al., 2016), and aligned-embeddings MUSE (Conneau et al., 2017). A standard method to utilize monolingual embeddings is to concatenate or sum two or more embeddings from different languages (Trivedi et al., 2018). A more recent approach is to apply an attention mechanism to merge embeddings and form meta-embeddings (Winata et al., 2019a,b). Character-based embeddings have also been explored in the literature to address the out-of-vocabulary issues on word-embeddings (Winata et al., 2018b; Attia et al., 2018; Aguilar et al., 2021). Another approach is to train bilingual embeddings using real and synthetic CSW data (Pratapa et al., 2018b). In the speech domain, Lovenia et al. (2022) utilize wav2vec 2.0 (Baevski et al., 2020) as a starting model before fine-tuning.

Language Models Many pre-trained model approaches utilize multilingual LMs, such as mBERT or XLM-R to deal with CSW data (Khanuja et al., 2020b; Aguilar and Solorio, 2020; Pant and Dadu, 2020; Patwa et al., 2020; Winata et al., 2021a). These models are often fine-tuned with the downstream task or with CSW text to better adapt to the languages. Some downstream fine-tuning approaches use synthetic CSW data due to a lack of available datasets. Aguilar et al. (2021) propose a character-based subword module (char2subword) of the mBERT that learns the subword embedding that is suitable for modeling the noisy CSW text. Winata et al. (2021a) compare the performance of the multilingual LM versus the language-specific LM for CSW context. While XLM-R provides the best result, it is also computationally heavy. There needed to be more exploration of larger models.

We see that pre-trained LMs provide better empirical results on current benchmark tasks and enables an end-to-end approach. Therefore, one can theoretically work on CSW tasks without any linguistic understanding of the language, assuming the dataset for model finetuning is available. However, the downside is that there is little understanding of how and when the LMs would fail, thus we encourage more interpretability work on these LMs in CSW setting.

6 Recent Challenges and Future Direction

6.1 More Diverse Exploration on Code-Switching Styles and Languages

A handful of languages, such as Spanish-English, Hindi-English, or Chinese-English, dominate research and resource CSW. However, there are still many countries and cultures rich in the use of CSW, which is still under-represented in NLP research (Joshi et al., 2020; Aji et al., 2022; Yong et al., 2023), especially on different CSW variations. CSW style can vary in different regions of the world, and it would be interesting to gather more datasets on unexplored and unknown styles, which can be useful for further research and investigation on linguistics and NLP. Therefore, one future direction is to broaden the language scope of CSW research.

6.2 Datasets: Access and Sources

According to our findings, there are more than 60% of the datasets are private (see Table 4), and they are not released to the public. This eventually hampers the progress of CSW research, particularly in the results' reproducibility, credibility, and transparency. Moreover, many studies in the literature do not release the code to reproduce their work. Therefore, we encourage researchers who build a new corpus to release the datasets publicly. In addition, the fact that some researchers provide urls to download the data is also problematic due to the data attrition issue we raised earlier. Data attrition is bad for reproducibility, but it is also a waste of annotation efforts. Perhaps we should work on identifying alternative means to collect written CSW data in an ecologically valid manner.

6.3 Model Scaling

To the best of our knowledge, little work has been done on investigating how well the scaling law holds for code-mixed datasets. Winata et al. (2021a) demonstrate that the XLM-R-large model outperforms smaller pre-trained models on the NER and POS tasks in LinCE benchmark (Aguilar et al., 2020); however, the largest model in the study, which is the XLM-R-large model, only has 355 million parameters. Furthermore, they find that smaller models that combine word, subword, and character embeddings achieve comparable performance as mBERT while being faster in inference. Given the recent release of billion-sized large pre-trained multilingual models such as XGLM and BLOOM (Scao et al., 2022), we urge future research to study the scaling law and performance-compute trade-off in code-mixing tasks.

6.4 Zero-Shot and Few-Shot Exploration

The majority of pre-trained model approaches fine-tune their models to the downstream task. On the other hand, CSW data is considerably limited. With the rise of multilingual LMs, especially those that have been fine-tuned with prompt/instruction (Muennighoff et al., 2022; Ouyang et al., 2022; Winata et al., 2022), one direction is to see whether these LMs can handle CSW input in a zero-shot fashion. This work might also tie in with model scaling since larger models have shown better capability at zero-shot and few-shot settings (Winata et al., 2021b; Srivastava et al., 2022).

6.5 Robustness Evaluation

Since CSW is a widely common linguistic phenomenon, we argue that cross-lingual NLP benchmarks, such as XGLUE (Liang et al., 2020) and XTREME-R (Ruder et al., 2021), should incorporate linguistic CSW evaluation (Aguilar et al., 2020; Khanuja et al., 2020b). The reasons are that CSW is a cognitive ability that multilingual human speakers can perform with ease (Beatty-Martínez et al., 2020). CSW evaluation examines the robustness of multilingual LMs in learning cross-lingual alignment of representations (Conneau et al., 2020; Libovický et al., 2020; Pires et al., 2019; Adilazuarda et al., 2022). On the other hand, catastrophic forgetting is observed in pre-trained models (Shah et al., 2020), and human speakers (Hirvonen and Lauttamus, 2000; Du Bois, 2009, known as language attrition) in a CSW environment. We argue that finetuning LMs on code-mixed data is a form of *continual learning* to produce a more generalized multilingual LM. Thus, we encourage CSW research to report the performance of finetuned models on both CSW and monolingual texts.

6.6 Task Diversity

We encourage creating reasoning-based tasks for CSW texts for two reasons. First, code-mixed datasets for tasks such as NLI, coreference resolution, and question-answering are much fewer in comparison to tasks such as sentiment analysis, parts-of-speech tagging, and named-entity recognition. Second, comprehension tasks with the CSW text present more processing costs for human readers (Bosma and Pablos, 2020).

6.7 Conversational Agents

There has been a recent focus on developing conversational agents with LMs such as ChatGPT,[5] Whisper (Radford et al., 2022), SLAM (Bapna et al., 2021), mSLAM (Bapna et al., 2022). We recommend incorporating the capability of synthesizing code-mixed data in human-machine dialogue, as CS is a prevalent communication style among multilingual speakers (Ahn et al., 2020), and humans prefer chatbots with such capability (Bawa et al., 2020).

6.8 Automatic Evaluation for Generation

With the rise of pre-trained models, generative tasks gained more popularity. However, when generating CSW data, most work used human evaluation for measuring quality of the generated data. Alternative automatic methods for CS text based on word-frequency and temporal distribution are commonly used (Guzmán et al., 2017; Mave et al., 2018), but we believe there is still much room for improvement in this respect. One possible future direction is to align the evaluation metrics to human judgement of quality (Hamed et al., 2022) where we can assess separately the "faithfulness" of the resulting CSW data from other desired properties of language generation. Other nuances here are related to the intricacy of CSW patterns, where ideally the model would mimic the CSW style of the intended users.

[5] https://openai.com/blog/chatgpt/

7 Conclusion

We present a comprehensive systematic survey on code-switching research in natural language processing to explore the progress of the past decades and understand the existing challenges and tasks in the literature. We summarize the trends and findings and conclude with a discussion for future direction and open questions for further investigation. We hope this survey can encourage and lead NLP researchers in a better direction on code-switching research.

Limitations

The numbers in this survey are limited to papers published in the ACL Anthology and ISCA Proceedings. However, we also included papers as related work from other resources if they are publicly available and accessible. In addition, the category in the survey does not include the code-switching type (i.e., intra-sentential, inter-sentential, etc.) since some papers do not provide such information.

Ethics Statement

We use publicly available data in our survey with permissive licenses. No potential ethical issues in this work.

Acknowledgements

Thanks to Igor Malioutov for the insightful discussion on the paper.

References

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020. Toward micro-dialect identification in diaglossic and code-switched environments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5855–5876.

Heike Adel, Katrin Kirchhoff, Dominic Telaar, Ngoc Thang Vu, Tim Schlippe, and Tanja Schultz. 2014a. Features for factored language models for code-switching speech. In *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 32–38.

Heike Adel, Katrin Kirchhoff, Ngoc Thang Vu, Dominic Telaar, and Tanja Schultz. 2014b. Comparing approaches to convert recurrent neural networks into backoff language models for efficient decoding. In *Proc. Interspeech 2014*, pages 651–655.

Heike Adel, Dominic Telaar, Ngoc Thang Vu, Katrin Kirchhoff, and Tanja Schultz. 2014c. Combining recurrent neural networks and factored language models during decoding of code-switching speech. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Heike Adel, Ngoc Thang Vu, and Tanja Schultz. 2013. Combination of recurrent neural networks and factored language models for code-switching language modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–211.

Muhammad Farid Adilazuarda, Samuel Cahyawijaya, Genta Indra Winata, Pascale Fung, and Ayu Purwarianti. 2022. Indorobusta: Towards robustness against diverse code-mixed indonesian local languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 25–34.

Wafia Adouane and Jean-Philippe Bernardy. 2020. When is multi-task learning beneficial for low-resource noisy code-switched user-generated algerian texts? In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 17–25.

Wafia Adouane, Jean-Philippe Bernardy, and Simon Dobnik. 2018. Improving neural network performance by injecting background knowledge: Detecting code-switching and borrowing in algerian texts. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 20–28.

Wafia Adouane, Samia Touileb, and Jean-Philippe Bernardy. 2020. Identifying sentiments in algerian code-switched user-generated comments. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2698–2705.

Laksh Advani, Clement Lu, and Suraj Maharjan. 2020. C1 at semeval-2020 task 9: Sentimix: Sentiment analysis for code-mixed social media text using feature engineering. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1227–1232.

Kaustubh Agarwal and Rhythm Narula. 2021. Humor generation and detection in code-mixed hindi-english. In *Proceedings of the Student Research Workshop Associated with RANLP 2021*, pages 1–6.

Vibhav Agarwal, Pooja Rao, and Dinesh Babu Jayagopi. 2021. Towards code-mixed hinglish dialogue generation. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 271–280.

Akshita Aggarwal, Anshul Wadhawan, Anshima Chaudhary, and Kavita Maurya. 2020. "did you really mean what you said?": Sarcasm detection in hindi-english code-mixed data using bilingual word embeddings. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 7–15.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018. Named entity recognition on code-switched data: Overview of the calcs 2018 shared task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. Lince: A centralized benchmark for linguistic code-switching evaluation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1803–1813.

Gustavo Aguilar, Bryan McCann, Tong Niu, Nazneen Rajani, Nitish Shirish Keskar, and Thamar Solorio. 2021. Char2subword: Extending the subword embedding space using robust character compositionality. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1640–1651.

Gustavo Aguilar and Thamar Solorio. 2020. From english to code-switching: Transfer learning with strong morphological clues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8033–8044.

Maia Aguirre, Laura García-Sardiña, Manex Serras, Ariane Méndez, and Jacobo López. 2022. Basco: An annotated basque-spanish code-switching corpus for natural language understanding. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3158–3163.

Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan W Black. 2020. What code-switching strategies are effective in dialog systems? In *Proceedings of the Society for Computation in Linguistics 2020*, pages 254–264.

Alham Aji, Genta Indra Winata, Fajri Koto, Samuel Cahyawijaya, Ade Romadhony, Rahmad Mahendra, Kemal Kurniawan, David Moeljadi, Radityo Eko Prasojo, Timothy Baldwin, et al. 2022. One country, 700+ languages: Nlp challenges for underrepresented languages and dialects in indonesia. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7226–7249.

Mohamed Al-Badrashiny and Mona Diab. 2016. The george washington university system for the code-switching workshop shared task 2016. In *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pages 108–111.

Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. 2016. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107.

Ahmed Ali, Shammur Absar Chowdhury, Amir Hussein, and Yasser Hifny. 2021. Arabic code-switching speech recognition using monolingual data. In *Proc. Interspeech 2021*, pages 3475–3479.

Djegdjiga Amazouz, Martine Adda-Decker, and Lori Lamel. 2017. Addressing code-switching in french/algerian arabic speech. In *Proc. Interspeech 2017*, pages 62–66.

Saadullah Amin, Noon Pokaratsiri Goldstein, Morgan Wixted, Alejandro Garcia-Rudolph, Catalina Martínez-Costa, and Günter Neumann. 2022. Few-shot cross-lingual transfer for coarse-grained de-identification of code-mixed clinical texts. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 200–211.

Judith Jeyafreeda Andrew. 2021. Judithjeyafreedaandrew@ dravidianlangtech-eacl2021: offensive language detection for dravidian code-mixed youtube comments. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 169–174.

Jason Angel, Segun Taofeek Aroyehun, Antonio Tamayo, and Alexander Gelbukh. 2020. Nlp-cic at semeval-2020 task 9: Analysing sentiment in code-switching language using a simple deep-learning classifier. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 957–962.

Ansen Antony, Sumanth Reddy Kota, Akhilesh Lade, Spoorthy V, and Shashidhar G. Koolagudi. 2022. An improved transformer transducer architecture for hindi-english code switched speech recognition. In *Proc. Interspeech 2022*, pages 3123–3127.

Lavinia Aparaschivei, Andrei Palihovici, and Daniela Gîfu. 2020. Fii-uaic at semeval-2020 task 9: Sentiment analysis for code-mixed social media text using cnn. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 928–933.

Ramakrishna Appicharla, Kamal Kumar Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2021. Iitp-mt at calcs2021: English to hinglish neural machine translation using unsupervised synthetic code-mixed parallel corpus. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 31–35.

Dian Arianto and Indra Budi. 2020. Aspect-based sentiment analysis on indonesia's tourism destinations based on google maps user code-mixed reviews (study case: Borobudur and prambanan temples). In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 359–367.

Kavita Asnani and Jyoti Pawar. 2016. Use of semantic knowledge base for enhancement of coherence of code-mixed topic-based aspect clusters. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 259–266.

Mohammed Attia, Younes Samih, and Wolfgang Maier. 2018. Ghht at calcs 2018: Named entity recognition for dialectal arabic using neural networks. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 98–102.

Leonardo Badino, Claudia Barolo, and Silvia Quazza. 2004. A general approach to tts reading of mixed-language texts. In *Proc. Interspeech 2004*, pages 849–852.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Mohamed Balabel, Injy Hamed, Slim Abdennadher, Ngoc Thang Vu, and Özlem Çetinoğlu. 2020. Cairo student code-switch (cscs) corpus: An annotated egyptian arabic-english corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3973–3977.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching*, pages 116–126.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Association for Computational Linguistics*.

Fazlourrahman Balouchzahi, BK Aparna, and HL Shashirekha. 2021. Mucs@ dravidianlangtech-eacl2021: Cooli-code-mixing offensive language identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329.

Fazlourrahman Balouchzahi and HL Shashirekha. 2021. La-saco: A study of learning approaches for sentiments analysis incode-mixing texts. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 109–118.

Somnath Banerjee, Sahar Ghannay, Sophie Rosset, Anne Vilnat, and Paolo Rosso. 2020. Limsi_upv at semeval-2020 task 9: Recurrent convolutional neural network for code-mixed sentiment analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1281–1287.

Suman Banerjee, Nikita Moghe, Siddhartha Arora, and Mitesh M Khapra. 2018. A dataset for building code-mixed goal oriented conversation systems. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3766–3780.

Neetika Bansal, Vishal Goyal, and Simpel Rani. 2020a. Language identification and normalization of code mixed english and punjabi text. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON): System Demonstrations*, pages 30–31.

Shubham Bansal, Arijit Mukherjee, Sandeepkumar Satpal, and Rupeshkumar Mehta. 2020b. On improving code mixed speech synthesis with mixlingual grapheme-to-phoneme model. In *Proc. Interspeech 2020*, pages 2957–2961.

Srijan Bansal, Vishal Garimella, Ayush Suhane, Jasabanta Patro, and Animesh Mukherjee. 2020c. Code-switching patterns can be an effective route to improve performance of downstream nlp applications: A case study of humour, sarcasm and hate speech detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1018–1023.

Ankur Bapna, Colin Cherry, Yu Zhang, Ye Jia, Melvin Johnson, Yong Cheng, Simran Khanuja, Jason Riesa, and Alexis Conneau. 2022. mslam: Massively multilingual joint pre-training for speech and text. *arXiv preprint arXiv:2202.01374*.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

Kfir Bar and Nachum Dershowitz. 2014. The tel aviv university system for the code-switching workshop shared task. In *Proceedings of the first workshop on computational approaches to code switching*, pages 139–143.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014a. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching*, pages 13–23.

Utsab Barman, Joachim Wagner, Grzegorz Chrupała, and Jennifer Foster. 2014b. Dcu-uvt: Word-level language classification with code-mixed data. In *Proceedings of the first workshop on computational approaches to code switching*, pages 127–132.

Utsab Barman, Joachim Wagner, and Jennifer Foster. 2016. Part-of-speech tagging of code-mixed social media content: Pipeline, stacking and joint modelling. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 30–39.

Subhra Jyoti Baroi, Nivedita Singh, Ringki Das, and Thoudam Doren Singh. 2020. Nits-hinglish-sentimix at semeval-2020 task 9: Sentiment analysis for code-mixed social media text using an ensemble model. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1298–1303.

Arup Baruah, Kaushik Das, Ferdous Barbhuiya, and Kuntal Dey. 2020. Iiitg-adbu at semeval-2020 task 12: Comparison of bert and bilstm in detecting offensive language. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1562–1568.

Anshul Bawa, Monojit Choudhury, and Kalika Bali. 2018. Accommodation of conversational code-choice. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 82–91.

Anshul Bawa, Pranav Khadpe, Pratik Joshi, Kalika Bali, and Monojit Choudhury. 2020. Do multilingual users prefer chat-bots that code-mix? let's nudge and find out! *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–23.

Anne L Beatty-Martínez, Christian A Navarro-Torres, and Paola E Dussias. 2020. Codeswitching: A bilingual toolkit for opportunistic speech planning. *Frontiers in Psychology*, page 1699.

Rafiya Begum, Kalika Bali, Monojit Choudhury, Koustav Rudra, and Niloy Ganguly. 2016. Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650.

Hedi M Belazi, Edward J Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic inquiry*, pages 221–237.

B Bharathi et al. 2021. Ssncse_nlp@ dravidianlangtech-eacl2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318.

Irshad Bhat, Riyaz Ahmad Bhat, Manish Shrivastava, and Dipti Misra Sharma. 2017. Joining hands: Exploiting monolingual treebanks for parsing of code-mixing data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 324–330.

Shishir Bhattacharja. 2010. Benglish verbs: A case of code-mixing in bengali. In *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, pages 75–84.

Shankar Biradar and Sunil Saumya. 2022. Iiitdwd@ tamilnlp-acl2022: Transformer-based approach to classify abusive content in dravidian code-mixed text. In *Proceedings of the second workshop on speech and language technologies for Dravidian languages*, pages 100–104.

Astik Biswas, Febe De Wet, Thomas Niesler, et al. 2020. Semi-supervised acoustic and language model training for english-isizulu code-switched speech recognition. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 52–56.

Astik Biswas, Febe de Wet, Ewald van der Westhuizen, Emre Yilmaz, and Thomas Niesler. 2018a. Multilingual neural network acoustic modelling for asr of under-resourced english-isizulu code-switched speech. In *INTERSPEECH*, pages 2603–2607.

Astik Biswas, Ewald van der Westhuizen, Thomas Niesler, and Febe de Wet. 2018b. Improving asr for code-switched speech in under-resourced languages using out-of-domain data. In *SLTU*, pages 122–126.

Astik Biswas, Emre Yılmaz, Febe de Wet, Ewald van der Westhuizen, and Thomas Niesler. 2019. Semi-supervised acoustic model training for five-lingual code-switched asr. *Proc. Interspeech 2019*, pages 3745–3749.

Evelyn Bosma and Leticia Pablos. 2020. Switching direction modulates the engagement of cognitive control in bilingual reading comprehension: An erp study. *Journal of Neurolinguistics*, 55:100894.

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, frisian-dutch data. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 50–58.

Barbara Bullock, Wally Guzmán, Jacqueline Serigos, Vivek Sharath, and Almeida Jacqueline Toribio. 2018a. Predicting the presence of a matrix language in code-switching. In *Proceedings of the third workshop on computational approaches to linguistic code-switching*, pages 68–75.

Barbara E. Bullock, Gualberto Guzmán, Jacqueline Serigos, and Almeida Jacqueline Toribio. 2018b. Should code-switching models be asymmetric? In *Proc. Interspeech 2018*, pages 2534–2538.

Jesús Calvillo, Le Fang, Jeremy Cole, and David Reitter. 2020. Surprisal predicts code-switching in chinese-english bilingual text. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4029–4039.

Marguerite Cameron. 2020. Voice onset time en code-switching anglais-français: une étude des occlusives sourdes en début de mot (voice onset time in english-french code-switching: a study of word-initial voiceless stop consonants). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1: Journées d'Études sur la Parole*, pages 54–63.

Houwei Cao, P. C. Ching, and Tan Lee. 2009. Effects of language mixing for automatic recognition of cantonese-english code-mixing utterances. In *Proc. Interspeech 2009*, pages 3011–3014.

Marine Carpuat. 2014. Mixed language and code-switching in the canadian hansard. In *Proceedings of the first workshop on computational approaches to code switching*, pages 107–115.

Özlem Çetinoğlu. 2016. A turkish-german code-switching corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4215–4220.

Özlem Çetinoğlu and Çağrı Çöltekin. 2019. Challenges of annotating a code-switching treebank. In *Proceedings of the 18th international workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 1–11.

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.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John Philip McCrae. 2020a. A sentiment analysis dataset for code-mixed malayalam-english. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 177–184.

Bharathi Raja Chakravarthi, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John Philip McCrae. 2020b. Corpus creation for sentiment analysis in code-mixed tamil-english text. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 202–210.

Sharanya Chakravarthy, Anjana Umapathy, and Alan W Black. 2020. Detecting entailment in code-mixed hindi-english conversations. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020)*, pages 165–170.

Joyce YC Chan, Houwei Cao, PC Ching, and Tan Lee. 2009. Automatic recognition of cantonese-english code-mixing speech. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 14, Number 3, September 2009*.

Joyce YC Chan, PC Ching, and Tan Lee. 2005. Development of a cantonese-english code-mixing speech corpus. In *Ninth European conference on speech communication and technology*.

Joyce YC Chan, PC Ching, Tan Lee, and Houwei Cao. 2006. Automatic speech recognition of cantonese-english code-mixing utterances. In *Ninth International Conference on Spoken Language Processing*.

Joyce YC Chan, PC Ching, Tan Lee, and Helen M Meng. 2004. Detection of language boundary in code-switching utterances by bi-phone probabilities. In *2004 International Symposium on Chinese Spoken Language Processing*, pages 293–296. IEEE.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016a. Columbia-jadavpur submission for emnlp 2016 code-switching workshop shared task: System description. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 112–115.

Arunavha Chanda, Dipankar Das, and Chandan Mazumdar. 2016b. Unraveling the english-bengali code-mixing phenomenon. In *Proceedings of the second workshop on computational approaches to code switching*, pages 80–89.

Khyathi Chandu, Ekaterina Loginova, Vishal Gupta, Josef van Genabith, Günter Neumann, Manoj Chinnakotla, Eric Nyberg, and Alan W Black. 2019. Code-mixed question answering challenge: Crowdsourcing data and techniques. In *Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 29–38. Association for Computational Linguistics (ACL).

Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W Black. 2018. Language informed modeling of code-switched text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 92–97.

Khyathi Raghavi Chandu and Alan W. Black. 2020. Style variation as a vantage point for code-switching. In *Proc. Interspeech 2020*, pages 4761–4765.

Khyathi Raghavi Chandu, SaiKrishna Rallabandi, Sunayana Sitaram, and Alan W. Black. 2017. Speech synthesis for mixed-language navigation instructions. In *Proc. Interspeech 2017*, pages 57–61.

Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee. 2019. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. *Proc. Interspeech 2019*, pages 554–558.

Arindam Chatterjere, Vineeth Guptha, Parul Chopra, and Amitava Das. 2020. Minority positive sampling for switching points-an anecdote for the code-mixing language modeling. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6228–6236.

Sik Feng Cheong, Hai Leong Chieu, and Jing Lim. 2021. Intrinsic evaluation of language models for code-switching. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 81–86.

Dhivya Chinnappa. 2021. dhivya-hope-detection@ lt-edi-eacl2021: multilingual hope speech detection for code-mixed and transliterated texts. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 73–78.

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and Monojit Choudhury. 2014. Word-level language identification using crf: Code-switching shared task report of msr india system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 73–79.

Won Ik Cho, Seok Min Kim, and Nam Soo Kim. 2020. Towards an efficient code-mixed grapheme-to-phoneme conversion in an agglutinative language: A case study on to-korean transliteration. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 65–70.

Parul Chopra, Sai Krishna Rallabandi, Alan W Black, and Khyathi Raghavi Chandu. 2021. Switch point biased self-training: Re-purposing pretrained models for code-switching. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4389–4397.

Monojit Choudhury, Kalika Bali, Sunayana Sitaram, and Ashutosh Baheti. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 65–74.

Shammur Absar Chowdhury, Amir Hussein, Ahmed Abdelali, and Ahmed Ali. 2021. Towards one model to rule all: Multilingual strategy for dialectal code-switching arabic asr. In *Proc. Interspeech 2021*, pages 2466–2470.

Chyng-Leei Chu, Dau-cheng Lyu, and Ren-yuan Lyu. 2007. Language identification on code-switching speech. In *Proceedings of ROCLING*.

Daniel Claeser, Samantha Kent, and Dennis Felske. 2018. Multilingual named entity recognition on spanish-english code-switched tweets using support vector machines. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 132–137.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

Bhargav Dave, Shripad Bhat, and Prasenjit Majumder. 2021. Irnlp_daiict@ dravidianlangtech-eacl2021: offensive language identification in dravidian languages using tf-idf char n-grams and muril. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 266–269.

Frances Adriana Laureano De Leon, Florimond Guéniat, and Harish Tayyar Madabushi. 2020. Cs-embed at semeval-2020 task 9: The effectiveness of code-switched word embeddings for sentiment analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 922–927.

Anik Dey and Pascale Fung. 2014. A hindi-english code-switching corpus. In *LREC*, pages 2410–2413.

Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Nada Almarwani, and Mohamed Al-Badrashiny. 2016. Creating a large multi-layered representational repository of linguistic code switched arabic data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4228–4235.

Mona Diab and Ankit Kamboj. 2011. Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for hindi english code switched data: A pilot annotation. In *Proceedings of the 9th Workshop on Asian Language Resources*, pages 36–40.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. Mucs 2021: Multilingual and code-switching asr challenges for low resource indian languages. In *Proc. Interspeech 2021*, pages 2446–2450.

Amazouz Djegdjiga, Martine Adda-Decker, and Lori Lamel. 2018. The french-algerian code-switching triggered audio corpus (facst). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

A Seza Doğruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666.

Manoel Verissimo dos Santos Neto, Ayrton Amaral, Nádia Silva, and Anderson da Silva Soares. 2020. Deep learning brasil-nlp at semeval-2020 task 9: sentiment analysis of code-mixed tweets using ensemble of language models. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1233–1238.

Suman Dowlagar and Radhika Mamidi. 2021a. Gated convolutional sequence to sequence based learning for english-hingilsh code-switched machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 26–30.

Suman Dowlagar and Radhika Mamidi. 2021b. Graph convolutional networks with multi-headed attention for code-mixed sentiment analysis. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 65–72.

Suman Dowlagar and Radhika Mamidi. 2021c. Offlangone@ dravidianlangtech-eacl2021: Transformers with the class balanced loss for offensive language identification in dravidian code-mixed text. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 154–159.

Suman Dowlagar and Radhika Mamidi. 2022. Cmnerone at semeval-2022 task 11: Code-mixed named entity recognition by leveraging multilingual data. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1556–1561.

Inke Du Bois. 2009. Language attrition and code-switching among us americans in germany. *Stellenbosch papers in linguistics PLUS*, 39:1–16.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip R Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 379–389.

Aparna Dutta. 2022. Word-level language identification using subword embeddings for code-mixed bangla-english social media data. In *Proceedings of the Workshop on Dataset Creation for Lower-Resourced Languages within the 13th Language Resources and Evaluation Conference*, pages 76–82.

Mahmoud El-Haj, Paul Rayson, and Mariam Aboelezz. 2018. Arabic dialect identification in the context of bivalency and code-switching. In *Proceedings of the 11th International Conference on Language Resources and Evaluation, Miyazaki, Japan.*, pages 3622–3627. European Language Resources Association.

Abdellah El Mekki, Abdelkader El Mahdaouy, Mohammed Akallouch, Ismail Berrada, and Ahmed Khoumsi. 2022. Um6p-cs at semeval-2022 task 11: Enhancing multilingual and code-mixed complex named entity recognition via pseudo labels using multilingual transformer. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1511–1517.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2014. Aida: Identifying code switching in informal arabic text. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 94–101.

Heba Elfardy and Mona Diab. 2012. Token level identification of linguistic code switching. In *Proceedings of COLING 2012: posters*, pages 287–296.

AbdelRahim Elmadany, Muhammad Abdul-Mageed, et al. 2021. Investigating code-mixed modern standard arabic-egyptian to english machine translation. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 56–64.

Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash, and Owen Rambow. 2014. Foreign words and the automatic processing of arabic social media text written in roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 1–12.

Guokang Fu and Liqin Shen. 2000. Model distance and it's application on mixed language speech recognition system. *ISCSLP'2000*.

Ruibo Fu, Jianhua Tao, Zhengqi Wen, Jiangyan Yi, Chunyu Qiang, and Tao Wang. 2020. Dynamic soft windowing and language dependent style token for code-switching end-to-end speech synthesis. In *Proc. Interspeech 2020*, pages 2937–2941.

Pascale Fung, Xiaohu Liu, and Chi-Shun Cheung. 1999. Mixed language query disambiguation. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 333–340.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Sreeram Ganji and Rohit Sinha. 2018. A novel approach for effective recognition of the code-switched data on monolingual language model. In *Proc. Interspeech 2018*, pages 1953–1957.

Yingying Gao, Junlan Feng, Ying Liu, Leijing Hou, Xin Pan, and Yong Ma. 2019. Code-switching sentence generation by bert and generative adversarial networks. In *Proc. Interspeech 2019*, pages 3525–3529.

Avishek Garain, Sainik Mahata, and Dipankar Das. 2020. Junlp at semeval-2020 task 9: Sentiment analysis of hindi-english code mixed data using grid search cross validation. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1276–1280.

Ayush Garg, Sammed Kagi, Vivek Srivastava, and Mayank Singh. 2021. Mipe: A metric independent pipeline for effective code-mixed nlg evaluation. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 123–132.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018a. Code-switched language models using dual rnns and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3078–3083.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018b. Dual language models for code switched speech recognition. In *Proc. Interspeech 2018*, pages 2598–2602.

Akash Kumar Gautam. 2022. Leveraging sub label dependencies in code mixed indian languages for part-of-speech tagging using conditional random fields. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 13–17.

Devansh Gautam, Kshitij Gupta, and Manish Shrivastava. 2021a. Translate and classify: Improving sequence level classification for english-hindi code-mixed data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 15–25.

Devansh Gautam, Prashant Kodali, Kshitij Gupta, Anmol Goel, Manish Shrivastava, and Ponnurangam Kumaraguru. 2021b. Comet: Towards code-mixed translation using parallel monolingual sentences. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 47–55.

Parvathy Geetha, Khyathi Chandu, and Alan W Black. 2018. Tackling code-switched ner: Participation of cmu. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 126–131.

Souvick Ghosh, Satanu Ghosh, and Dipankar Das. 2016. Part-of-speech tagging of code-mixed social media text. In *Proceedings of the second workshop on computational approaches to code switching*, pages 90–97.

Urmi Ghosh, Dipti Misra Sharma, and Simran Khanuja. 2019. Dependency parser for bengali-english code-mixed data enhanced with a synthetic treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 91–99.

Oluwapelumi Giwa and Marelie H. Davel. 2014. Language identification of individual words with joint sequence models. In *Proc. Interspeech 2014*, pages 1400–1404.

Hila Gonen and Yoav Goldberg. 2019. Language modeling for code-switching: Evaluation, integration of monolingual data, and discriminative training. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4175–4185.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Vinay Gopalan and Mark Hopkins. 2020. Reed at semeval-2020 task 9: Fine-tuning and bag-of-words approaches to code-mixed sentiment analysis. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1304–1309.

Koustava Goswami, Priya Rani, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. Uld@ nuig at semeval-2020 task 9: Generative morphemes with an attention model for sentiment analysis in code-mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 968–974.

Wen-Tao Gu, Tan Lee, and P. C. Ching. 2008. Prosodic variation in cantonese-english code-mixed speech. In *Proc. International Symposium on Chinese Spoken Language Processing*, pages 342–345.

Sunil Gundapu and Radhika Mamidi. 2018. Word level language identification in english telugu code mixed data. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*.

Sunil Gundapu and Radhika Mamidi. 2020. Gundapusunil at semeval-2020 task 9: Syntactic semantic lstm architecture for sentiment analysis of code-mixed data. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1247–1252.

Pengcheng Guo, Haihua Xu, Lei Xie, and Eng Siong Chng. 2018. Study of semi-supervised approaches to improving english-mandarin code-switching speech recognition. In *Proc. Interspeech 2018*, pages 1928–1932.

Abhirut Gupta, Aditya Vavre, and Sunita Sarawagi. 2021a. Training data augmentation for code-mixed translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5760–5766.

Akshat Gupta, Sargam Menghani, Sai Krishna Rallabandi, and Alan W Black. 2021b. Unsupervised self-training for sentiment analysis of code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–112.

Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black. 2021c. Task-specific pre-training and cross lingual transfer for sentiment analysis in dravidian code-switched languages. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 73–79.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2018a. A deep neural network based approach for entity extraction in code-mixed indian social media

text. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280.

Deepak Gupta, Ankit Lamba, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Opinion mining in a code-mixed environment: A case study with government portals. In *Proceedings of the 13th International Conference on Natural Language Processing*, pages 249–258.

Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2018b. Uncovering code-mixed challenges: A framework for linguistically driven question generation and neural based question answering. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 119–130.

Vishal Gupta, Manoj Chinnakotla, and Manish Shrivastava. 2018c. Transliteration better than translation? answering code-mixed questions over a knowledge base. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 39–50.

Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *INTERSPEECH*, pages 67–71.

Gualberto A Guzman, Jacqueline Serigos, Barbara Bullock, and Almeida Jacqueline Toribio. 2016. Simple tools for exploring variation in code-switching for linguists. In *Proceedings of the second workshop on computational approaches to code switching*, pages 12–20.

Gualberto Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Proc. Interspeech 2017*, pages 67–71.

Injy Hamed, Mohamed Elmahdy, and Slim Abdennadher. 2018. Collection and analysis of code-switch egyptian arabic-english speech corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Injy Hamed, Amir Hussein, Oumnia Chellah, Shammur Chowdhury, Hamdy Mubarak, Sunayana Sitaram, Nizar Habash, and Ahmed Ali. 2022. Benchmarking evaluation metrics for code-switching automatic speech recognition. *arXiv preprint arXiv:2211.16319*.

Injy Hamed, Ngoc Thang Vu, and Slim Abdennadher. 2020. Arzen: A speech corpus for code-switched egyptian arabic-english. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4237–4246.

Silvana Hartmann, Monojit Choudhury, and Kalika Bali. 2018. An integrated representation of linguistic and social functions of code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ji He, Yao Qian, Frank K Soong, and Sheng Zhao. 2012. Turning a monolingual speaker into multilingual for a mixed-language tts. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha, and Bharathi Raja Chakravarthi. 2022. Corpus creation for sentiment analysis in code-mixed tulu text. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 33–40.

Megan Herrera, Ankit Aich, and Natalie Parde. 2022. Tweettaglish: A dataset for investigating tagalog-english code-switching. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2090–2097.

Pekka Hirvonen and Timo Lauttamus. 2000. Code-switching and language attrition: Evidence from american finnish interview speech. *SKY journal of linguistics*, 13:47–74.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ lt-edi-eacl2021: Multilingual code-mixed hope speech detection using cross-lingual representation learner. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 168–174.

Xinhui Hu, Qi Zhang, Lei Yang, Binbin Gu, and Xinkang Xu. 2020. Data augmentation for code-switch language modeling by fusing multiple text generation methods. In *Proc. Interspeech 2020*, pages 1062–1066.

Bo Huang and Yang Bai. 2021. hub at semeval-2021 task 7: Fusion of albert and word frequency information detecting and rating humor and offense. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1141–1145.

Fei Huang and Alexander Yates. 2014. Improving word alignment using linguistic code switching data. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1–9.

Dana-Maria Iliescu, Rasmus Grand, Sara Qirko, and Rob van der Goot. 2021. Much gracias: Semi-supervised code-switch detection for spanish-english: How far can we get? *NAACL 2021*, page 65.

David Imseng, Hervé Bourlard, and Mathew Magimai Doss. 2010. Towards mixed language speech recognition systems. In *Proc. Interspeech 2010*, pages 278–281.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A Smith. 2016. A neural model for language identification in code-switched tweets. In *Proceedings of The Second Workshop on Computational Approaches to Code Switching*, pages 60–64.

Devanshu Jain, Maria Kustikova, Mayank Darbari, Rishabh Gupta, and Stephen Mayhew. 2018. Simple features for strong performance on named entity recognition in code-switched twitter data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 103–109.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in code-switching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 87–93.

Anupam Jamatia, Björn Gambäck, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages. In *Proceedings of the international conference recent advances in natural language processing*, pages 239–248.

Florian Janke, Tongrui Li, Eric Rincón, Gualberto A Guzman, Barbara Bullock, and Almeida Jacqueline Toribio. 2018. The university of texas system submission for the code-switching workshop shared task 2018. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 120–125.

Soroush Javdan, Behrouz Minaei-Bidgoli, et al. 2020. Iust at semeval-2020 task 9: Sentiment analysis for code-mixed social media text using deep neural networks and linear baselines. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1270–1275.

Ganesh Jawahar, Muhammad Abdul-Mageed, VS Laks Lakshmanan, et al. 2021. Exploring text-to-text transformers for english to hinglish machine translation with synthetic code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 36–46.

Sai Muralidhar Jayanthi, Kavya Nerella, Khyathi Raghavi Chandu, and Alan W Black. 2021. Codemixednlp: An extensible and open nlp toolkit for code-mixing. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 113–118.

Harsh Jhamtani, Suleep Kumar Bhogi, and Vaskar Raychoudhury. 2014. Word-level language identification in bi-lingual code-switched texts. In *Proceedings of the 28th Pacific Asia Conference on language, information and computing*, pages 348–357.

Lars Johanson. 1999. The dynamics of code-copying in language encounters. *Language encounters across time and space*, 3762.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th international conference on advanced computing and communication systems (ICACCS)*, pages 136–141. IEEE.

Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2482–2491.

Aravind Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Coling 1982: Proceedings of the Ninth International Conference on Computational Linguistics*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

David Jurgens, Stefan Dimitrov, and Derek Ruths. 2014. Twitter users# codeswitch hashtags!# moltoimportante# wow. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 51–61.

Laurent Kevers. 2022. Coswid, a code switching identification method suitable for under-resourced languages. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 112–121.

Humair Raj Khan, Deepak Gupta, and Asif Ekbal. 2021. Towards developing a multilingual and code-mixed visual question answering system by knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1753–1767.

Ankush Khandelwal, Sahil Swami, Syed S Akhtar, and Manish Shrivastava. 2018. Humor detection in english-hindi code-mixed social media content: Corpus and baseline system. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Simran Khanuja, Sandipan Dandapat, Sunayana Sitaram, and Monojit Choudhury. 2020a. A new dataset for natural language inference from code-mixed conversations. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 9–16.

Simran Khanuja, Sandipan Dandapat, Anirudh Srinivasan, Sunayana Sitaram, and Monojit Choudhury. 2020b. Gluecos: An evaluation benchmark for code-switched nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3575–3585.

Yerbolat Khassanov, Haihua Xu, Van Tung Pham, Zhiping Zeng, Eng Siong Chng, Chongjia Ni, and Bin Ma. 2019. Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data. In *Proc. Interspeech 2019*, pages 2160–2164.

Levi King, Eric Baucom, Timur Gilmanov, Sandra Kübler, Daniel Whyatt, Wolfgang Maier, and Paul Rodrigues. 2014. The iucl+ system: Word-level language identification via extended markov models. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 102–106.

Ondřej Klejch, Electra Wallington, and Peter Bell. 2021. The cstr system for multilingual and code-switching asr challenges for low resource indian languages. In *Proc. Interspeech 2021*, pages 2881–2885.

Kate M. Knill, Linlin Wang, Yu Wang, Xixin Wu, and Mark J.F. Gales. 2020. Non-native children's automatic speech recognition: The interspeech 2020 shared task alta systems. In *Proc. Interspeech 2020*, pages 255–259.

Hiroaki Kojima and Kazuyo Tanaka. 2003. Mixed-lingual spoken word recognition by using vq codebook sequences of variable length segments. In *Eighth European Conference on Speech Communication and Technology*.

Jun Kong, Jin Wang, and Xuejie Zhang. 2020. Hpcc-ynu at semeval-2020 task 9: A bilingual vector gating mechanism for sentiment analysis of code-mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 940–945.

Ayush Kumar, Harsh Agarwal, Keshav Bansal, and Ashutosh Modi. 2020. Baksa at semeval-2020 task 9: Bolstering cnn with self-attention for sentiment analysis of code mixed text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1221–1226.

Mari Ganesh Kumar, Jom Kuriakose, Anand Thyagachandran, Arun Kumar A, Ashish Seth, Lodagala V.S.V. Durga Prasad, Saish Jaiswal, Anusha Prakash, and Hema A. Murthy. 2021. Dual script e2e framework for multilingual and code-switching asr. In *Proc. Interspeech 2021*, pages 2441–2445.

Ritesh Kumar, Aishwarya N Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of hindi-english code-mixed data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yash Kumar Lal, Vaibhav Kumar, Mrinal Dhar, Manish Shrivastava, and Philipp Koehn. 2019. De-mixing sentiment from code-mixed text. In *Proceedings of the 57th annual meeting of the association for computational linguistics: student research workshop*, pages 371–377.

Grandee Lee and Haizhou Li. 2020. Modeling code-switch languages using bilingual parallel corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 860–870.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019a. Linguistically motivated parallel data augmentation for code-switch language modeling. In *Interspeech*, pages 3730–3734.

Grandee Lee, Xianghu Yue, and Haizhou Li. 2019b. Linguistically motivated parallel data augmentation for code-switch language modeling. In *Proc. Interspeech 2019*, pages 3730–3734.

Chengfei Li, Shuhao Deng, Yaoping Wang, Guangjing Wang, Yaguang Gong, Changbin Chen, and Jinfeng Bai. 2022. Talcs: An open-source mandarin-english code-switching corpus and a speech recognition baseline. In *Proc. Interspeech 2022*, pages 1741–1745.

Chia-Yu Li and Ngoc Thang Vu. 2020. Improving code-switching language modeling with artificially generated texts using cycle-consistent adversarial networks. In *Proc. Interspeech 2020*, pages 1057–1061.

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*, pages 1671–1680.

Ying Li and Pascale Fung. 2013. Language modeling for mixed language speech recognition using weighted phrase extraction. In *Interspeech*, pages 2599–2603.

Ying Li and Pascale Fung. 2014. Language modeling with functional head constraint for code switching speech recognition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 907–916.

Ying Li, Yue Yu, and Pascale Fung. 2012. A mandarin-english code-switching corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2515–2519.

Zichao Li. 2021. Codewithzichao@ dravidianlangtech-eacl2021: Exploring multimodal transformers for meme classification in tamil language. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 352–356.

Hui Liang, Yao Qian, and Frank K Soong. 2007. An hmm-based bilingual (mandarin-english) tts. *Proceedings of SSW6*.

Wei-Bin Liang, Chung-Hsien Wu, and Chun-Shan Hsu. 2013. Code-switching event detection based on delta-bic using phonetic eigenvoice models. In *Proc. Interspeech 2013*, pages 1487–1491.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018.

Jindřich Libovickỳ, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1663–1674.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. 2014. The cmu submission for the shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 80–86.

Hou-An Lin and Chia-Ping Chen. 2021. Exploiting low-resource code-switching data to mandarin-english speech recognition systems. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 81–86.

Wei-Ting Lin and Berlin Chen. 2020. Exploring disparate language model combination strategies for mandarin-english code-switching asr. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 346–358.

Hexin Liu, Leibny Paola García Perera, Xinyi Zhang, Justin Dauwels, Andy W.H. Khong, Sanjeev Khudanpur, and Suzy J. Styles. 2021. End-to-end language diarization for bilingual code-switching speech. In *Proc. Interspeech 2021*, pages 1489–1493.

Jiaxiang Liu, Xuyi Chen, Shikun Feng, Shuohuan Wang, Xuan Ouyang, Yu Sun, Zhengjie Huang, and Weiyue Su. 2020. Kk2018 at semeval-2020 task 9: Adversarial training for code-mixing sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 817–823.

Khaled Lounnas, Mourad Abbas, and Mohamed Lichouri. 2021. Towards phone number recognition for code switched algerian dialect. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)*, pages 290–294.

Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Peng Xu, Yan Xu, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. 2022. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7259–7268.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J Barezi, et al. 2021. Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. *arXiv preprint arXiv:2112.06223*.

Yizhou Lu, Mingkun Huang, Hao Li, Jiaqi Guo, and Yanmin Qian. 2020. Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts. In *Proc. Interspeech 2020*, pages 4766–4770.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Proc. Interspeech 2008*, pages 711–714.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010a. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Eleventh Annual Conference of the International Speech Communication Association*.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2010b. Seame: a mandarin-english code-switching speech corpus in south-east asia. In *Proc. Interspeech 2010*, pages 1986–1989.

Tetyana Lyudovyk and Valeriy Pylypenko. 2014. Code-switching speech recognition for closely related languages. In *Proc. 4th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2014)*, pages 188–193.

Yili Ma, Liang Zhao, and Jie Hao. 2020. Xlp at semeval-2020 task 9: Cross-lingual models with focal loss for sentiment analysis of code-mixing language. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 975–980.

Koena Ronny Mabokela, Madimetja Jonas Manamela, and Mabu Manaileng. 2014. Modeling code-switching speech on under-resourced languages for language identification. In *Spoken Language Technologies for Under-Resourced Languages*.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. 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 2005–2011.

Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2021. Sentiment classification of code-mixed tweets using bi-directional rnn and language tags. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 28–35.

Piyush Makhija, Ankit Srivastava, and Anuj Gupta. 2020. hinglishnorm-a corpus of hindi-english code mixed sentences for text normalization. In *Proceedings of the 28th International Conference on Computational Linguistics: Industry Track*, pages 136–145.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022a. Multiconer: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022b. Semeval-2022 task 11: Multilingual complex named entity recognition (multiconer). In *Proceedings of the 16th international workshop on semantic evaluation (SemEval-2022)*, pages 1412–1437.

Aditya Malte, Pratik Bhavsar, and Sushant Rathi. 2020. Team_swift at semeval-2020 task 9: Tiny data specialists through domain-specific pre-training on code-mixed data. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1310–1315.

Soumil Mandal and Karthick Nanmaran. 2018. Normalization of transliterated words in code-mixed data using seq2seq model & levenshtein distance. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 49–53.

Soumil Mandal and Anil Kumar Singh. 2018. Language identification in code-mixed data using multichannel neural networks and context capture. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 116–120.

Asrita Venkata Mandalam and Yashvardhan Sharma. 2021. Sentiment analysis of dravidian code mixed data. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 46–54.

Sreeja Manghat, Sreeram Manghat, and Tanja Schultz. 2020. Malayalam-english code-switched: Grapheme to phoneme system. In *Proc. Interspeech 2020*, pages 4133–4137.

Sreeram Manghat, Sreeja Manghat, and Tanja Schultz. 2022. Normalization of code-switched text for speech synthesis. In *Proc. Interspeech 2022*, pages 4297–4301.

J. C. Marcadet, V. Fischer, and C. Waast-Richard. 2005. A transformation-based learning approach to language identification for mixed-lingual text-to-speech synthesis. In *Proc. Interspeech 2005*, pages 2249–2252.

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language identification and analysis of code-switched social media text. In *Proceedings of the third workshop on computational approaches to linguistic code-switching*, pages 51–61.

Laiba Mehnaz, Debanjan Mahata, Rakesh Gosangi, Uma Sushmitha Gunturi, Riya Jain, Gauri Gupta, Amardeep Kumar, Isabelle G Lee, Anish Acharya, and Rajiv Shah. 2021. Gupshup: Summarizing open-domain code-switched conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6177–6192.

Elena Álvarez Mellado and Constantine Lignos. 2022. Borrowing or codeswitching? annotating for finer-grained distinctions in language mixing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3195–3201.

Gideon Mendels, Victor Soto, Aaron Jaech, and Julia Hirschberg. 2018. Collecting code-switched data from social media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Siddhartha Mukherjee, Vinuthkumar Prasan, Anish Nediyanchath, Manan Shah, and Nikhil Kumar. 2019. Robust deep learning based sentiment classification of code-mixed text. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 124–129.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022a. Kazakhtts2: Extending the open-source kazakh tts corpus with more data, speakers, and topics. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5404–5411.

Saida Mussakhojayeva, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022b. Ksc2: An industrial-scale open-source kazakh speech corpus. In *Proc. Interspeech 2022*, pages 1367–1371.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Carol Myers-Scotton. 2005. *Multiple voices: An introduction to bilingualism*. John Wiley & Sons.

Ravindra Nayak and Raviraj Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12.

Tomáš Nekvinda and Ondřej Dušek. 2020. One model, many languages: Meta-learning for multilingual text-to-speech. In *Proc. Interspeech 2020*, pages 2972–2976.

Li Nguyen and Christopher Bryant. 2020. Canvec-the canberra vietnamese-english code-switching natural speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4121–4129.

Thomas Niesler and Febe de Wet. 2008. Accent identification in the presence of code-mixing. In *Proc. The Speaker and Language Recognition Workshop (Odyssey 2008)*, page paper 27.

Thomas Niesler et al. 2018. A first south african corpus of multilingual code-switched soap opera speech. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

José Carlos Rosales Núñez and Guillaume Wisniewski. 2018. Analyse morpho-syntaxique en présence d'alternance codique (pos tagging of code switching). In *Actes de la Conférence TALN. Volume 1-Articles longs, articles courts de TALN*, pages 473–480.

Nathaniel Oco and Rachel Edita Roxas. 2012. Pattern matching refinements to dictionary-based code-switching point detection. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 229–236.

Daniela Oria and Akos Vetek. 2004. Multilingual e-mail text processing for speech synthesis. In *Proc. Interspeech 2004*, pages 841–844.

Alissa Ostapenko, Shuly Wintner, Melinda Fricke, and Yulia Tsvetkov. 2022. Speaker information can guide models to better inductive biases: A case study on predicting code-switching. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3853–3867.

Billian Khalayi Otundo and Martine Grice. 2022. Intonation in advice-giving in kenyan english and kiswahili. In *Proc. Speech Prosody 2022*, pages 150–154.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Şaziye Özateş, Arzucan Özgür, Tunga Güngör, and Özlem Çetinoğlu. 2022. Improving code-switching dependency parsing with semi-supervised auxiliary tasks. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1159–1171.

Şaziye Betül Özateş and Özlem Çetinoğlu. 2021. A language-aware approach to code-switched morphological tagging. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 72–83.

Daniel Palomino and José Ochoa-Luna. 2020. Palomino-ochoa at semeval-2020 task 9: Robust system based on transformer for code-mixed sentiment classification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 963–967.

Ayushi Pandey, Brij Mohan Lal Srivastava, and Suryakanth Gangashetty. 2017. Towards developing a phonetically balanced code-mixed speech corpus for hindi-english asr. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 95–101.

Ayushi Pandey, Brij Mohan Lal Srivastava, Rohit Kumar, Bhanu Teja Nellore, Kasi Sai Teja, and Suryakanth V Gangashetty. 2018. Phonetically balanced code-mixed speech corpus for hindi-english automatic speech recognition. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kartikey Pant and Tanvi Dadu. 2020. Towards code-switched classification exploiting constituent language resources. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 37–43.

Evangelos Papalexakis, Dong Nguyen, and A Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *First Workshop on Computational Approaches to Code Switching (EMNLP 2014)*, pages 42–50. Association for Computational Linguistics (ACL).

Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 565–577.

Apurva Parikh, Abhimanyu Singh Bisht, and Prasenjit Majumder. 2020. Irlab_daiict at semeval-2020 task 9: Machine learning and deep learning methods for sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1265–1269.

Dwija Parikh and Thamar Solorio. 2021. Normalization and back-transliteration for code-switched data. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 119–124.

Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas Pykl, Björn Gambäck, Tanmoy

Chakraborty, Thamar Solorio, and Amitava Das. 2020. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 774–790.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–679.

Beat Pfister and Harald Romsdorfer. 2003. Mixed-lingual text analysis for polyglot tts synthesis. In *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*, pages 2037–2040.

Akshata Phadte and Gaurish Thakkar. 2017. Towards normalising konkani-english code-mixed social media text. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 85–94.

Page Piccinini and Marc Garellek. 2014. Prosodic cues to monolingual versus code-switching sentences in english and spanish. In *Proc. Speech Prosody 2014*, pages 885–889.

Mario Piergallini, Rouzbeh Shirvani, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in swahili-english language data. In *Proceedings of the second workshop on computational approaches to code switching*, pages 21–29.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001.

Shana Poplack. 1978. *Syntactic structure and social function of code-switching*, volume 2. Centro de Estudios Puertorriqueños,[City University of New York].

Shana Poplack. 1980. Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching. *Linguistics*, 18:581–618.

Anusha Prakash, Anju Leela Thomas, S. Umesh, and Hema A Murthy. 2019. Building multilingual end-to-end speech synthesisers for indian languages. In *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, pages 194–199.

Archiki Prasad, Mohammad Ali Rehan, Shreya Pathak, and Preethi Jyothi. 2021. The effectiveness of intermediate-task training for code-switched natural language understanding. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 176–190.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018a. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1543–1553.

Adithya Pratapa and Monojit Choudhury. 2017. Quantitative characterization of code switching patterns in complex multi-party conversations: A case study on hindi movie scripts. In *Proceedings of the 14th International Conference on Natural Language Processing (ICON-2017)*, pages 75–84.

Adithya Pratapa and Monojit Choudhury. 2021. Comparing grammatical theories of code-mixing. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 158–167.

Adithya Pratapa, Monojit Choudhury, and Sunayana Sitaram. 2018b. Word embeddings for code-mixed language processing. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3067–3072.

Aman Priyanshu, Aleti Vardhan, Sudarshan Sivakumar, Supriti Vijay, and Nipuna Chhabra. 2021. "something something hota hai!" an explainable approach towards sentiment analysis on indian code-mixed data. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 437–444.

Yao Qian, Houwei Cao, and Frank K Soong. 2008. Hmm-based mixed-language (mandarin-english) speech synthesis. In *2008 6th International Symposium on Chinese Spoken Language Processing*, pages 1–4. IEEE.

Zimeng Qiu, Yiyuan Li, Xinjian Li, Florian Metze, and William M. Campbell. 2020. Towards context-aware end-to-end code-switching speech recognition. In *Proc. Interspeech 2020*, pages 4776–4780.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Tathagata Raha, Sainik Mahata, Dipankar Das, and Sivaji Bandyopadhyay. 2019. Development of pos tagger for english-bengali code-mixed data. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 143–149.

Ratnavel Rajalakshmi, Yashwant Reddy, and Lokesh Kumar. 2021. Dlrg@ dravidianlangtech-eacl2021: Transformer based approachfor offensive language identification on code-mixed tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 357–362.

SaiKrishna Rallabandi and Alan W. Black. 2017. On building mixed lingual speech synthesis systems. In *Proc. Interspeech 2017*, pages 52–56.

SaiKrishna Rallabandi and Alan W. Black. 2019. Variational attention using articulatory priors for generating code mixed speech using monolingual corpora. In *Proc. Interspeech 2019*, pages 3735–3739.

SaiKrishna Rallabandi, Sunayana Sitaram, and Alan W Black. 2018. Automatic detection of code-switching style from acoustics. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 76–81.

Vikram Ramanarayanan and David Suendermann-Oeft. 2017. Jee haan, i'd like both, por favor: Elicitation of a code-switched corpus of hindi–english and spanish–english human–machine dialog. In *Proc. Interspeech 2017*, pages 47–51.

Banothu Rambabu and Suryakanth V Gangashetty. 2018. Development of iiith hindi-english code mixed speech database. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 107–111.

Priya Rani, John Philip McCrae, and Theodorus Fransen. 2022. Mhe: Code-mixed corpora for similar language identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3425–3433.

Priya Rani, Shardul Suryawanshi, Koustava Goswami, Bharathi Raja Chakravarthi, Theodorus Fransen, and John Philip McCrae. 2020. A comparative study of different state-of-the-art hate speech detection methods in hindi-english code-mixed data. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 42–48.

Preeti Rao, Mugdha Pandya, Kamini Sabu, Kanhaiya Kumar, and Nandini Bondale. 2018. A study of lexical and prosodic cues to segmentation in a hindi-english code-switched discourse. In *Proc. Interspeech 2018*, pages 1918–1922.

Manikandan Ravikiran and Subbiah Annamalai. 2021. Dosa: Dravidian code-mixed offensive span identification dataset. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 10–17.

Manikandan Ravikiran and Bharathi Raja Chakravarthi. 2022. Zero-shot code-mixed offensive span identification through rationale extraction. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 240–247.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, S Sangeetha, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, Shankar Mahadevan, et al. 2022. Findings of the shared task on offensive span identification fromcode-mixed tamil-english comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 261–270.

Xiaolin Ren, Xin He, and Yaxin Zhang. 2005. Mandarin/english mixed-lingual name recognition for mobile phone. In *INTERSPEECH*, pages 3373–3376.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Shekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1971–1982.

Harald Romsdorfer and Beat Pfister. 2005. Phonetic labeling and segmentation of mixed-lingual prosody databases. In *Proc. Interspeech 2005*, pages 3281–3284.

Harald Romsdorfer and Beat Pfister. 2006. Character stream parsing of mixed-lingual text. In *Proc. Multilingual Language and Speech Processing (MULTILING 2006)*, page paper 021.

Mike Rosner and Paulseph-John Farrugia. 2007. A tagging algorithm for mixed language identification in a noisy domain. In *Proc. Interspeech 2007*, pages 190–193.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, et al. 2021. Xtreme-r: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245.

Caroline Sabty, Mohamed Islam, and Slim Abdennadher. 2020. Contextual embeddings for arabic-english code-switched data. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 215–225.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016a. Multilingual code-switching identification via lstm recurrent neural networks. In *Proceedings of the second workshop on computational approaches to code switching*, pages 50–59.

Younes Samih and Wolfgang Maier. 2016. An arabic-moroccan darija code-switched corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4170–4175.

Younes Samih, Wolfgang Maier, and Laura Kallmeyer. 2016b. Sawt: Sequence annotation web tool. In *Proceedings of the second workshop on computational approaches to code switching*, pages 65–70.

David Sankoff. 1998. The production of code-mixed discourse. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Sebastin Santy, Anirudh Srinivasan, and Monojit Choudhury. 2021. Bertologicomix: How does code-mixing interact with multilingual bert? In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 111–121.

Sunil Saumya, Abhinav Kumar, and Jyoti Prakash Singh. 2021. Offensive language identification in dravidian code mixed social media text. In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pages 36–45.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Royal Sequiera, Monojit Choudhury, and Kalika Bali. 2015. Pos tagging of hindi-english code mixed text from social media: Some machine learning experiments. In *Proceedings of the 12th international conference on natural language processing*, pages 237–246.

Sanket Shah, Basil Abraham, Sunayana Sitaram, Vikas Joshi, et al. 2020. Learning to recognize code-switched speech without forgetting monolingual speech recognition. *arXiv preprint arXiv:2006.00782*.

Sanket Shah, Pratik Joshi, Sebastin Santy, and Sunayana Sitaram. 2019. Cossat: Code-switched speech annotation tool. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 48–52.

Omar Sharif, Eftekhar Hossain, and Mohammed Moshiul Hoque. 2021. Nlp-cuet@ dravidianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 255–261.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti Misra Sharma. 2016. Shallow parsing pipeline-hindi-english code-mixed social media text. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1340–1345.

Yash Sharma, Basil Abraham, Karan Taneja, and Preethi Jyothi. 2020. Improving low resource code-switched asr using augmented code-switched tts. In *Proc. Interspeech 2020*, pages 4771–4775.

Zhijie Shen and Wu Guo. 2022. An improved deliberation network with text pre-training for code-switching automatic speech recognition. In *Proc. Interspeech 2022*, pages 3854–3858.

Rouzbeh Shirvani, Mario Piergallini, Gauri Shankar Gautam, and Mohamed Chouikha. 2016. The howard university system submission for the shared task in language identification in spanish-english codeswitching. In *Proceedings of the second workshop on computational approaches to code switching*, pages 116–120.

Philippa Shoemark, James Kirby, and Sharon Goldwater. 2018. Inducing a lexicon of sociolinguistic variables from code-mixed text. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 1–6.

Prajwol Shrestha. 2014. Incremental n-gram approach for language identification in code-switched text. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 133–138.

Prajwol Shrestha. 2016. Codeswitching detection via lexical features in conditional random fields. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 121–126.

Zhiwei Shuang, Shiyin Kang, Yong Qin, Lirong Dai, and Lianhong Cai. 2010. Hmm based tts for mixed language text. In *Eleventh Annual Conference of the International Speech Communication Association*.

Utpal Kumar Sikdar, Biswanath Barik, and Björn Gambäck. 2018. Named entity recognition on code-switched data using conditional random fields. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 115–119.

Utpal Kumar Sikdar and Björn Gambäck. 2016. Language identification in code-switched text using conditional random fields and babelnet. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 127–131.

Anand Singh and Tien-Ping Tan. 2018. Evaluating code-switched malay-english speech using time delay neural networks. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 197–200.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58.

Pranaydeep Singh and Els Lefever. 2020. Sentiment analysis for hinglish code-mixed tweets by means of cross-lingual word embeddings. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 45–51.

Sunayana Sitaram and Alan W Black. 2016. Speech synthesis of code-mixed text. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3422–3428.

Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W Black. 2019. A survey of code-switched speech and language processing. *arXiv preprint arXiv:1904.00784*.

Sunayana Sitaram, Sai Krishna Rallabandi, Shruti Rijhwani, and Alan W. Black. 2016. Experiments with cross-lingual systems for synthesis of code-mixed text. In *Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, pages 76–81.

Sunit Sivasankaran, Brij Mohan Lal Srivastava, Sunayana Sitaram, Kalika Bali, and Monojit Choudhury. 2018. Phone merging for code-switched speech recognition. In *Third Workshop on Computational Approaches to Linguistic Code-switching*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981.

Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1060.

Tongtong Song, Qiang Xu, Meng Ge, Longbiao Wang, Hao Shi, Yongjie Lv, Yuqin Lin, and Jianwu Dang. 2022. Language-specific characteristic assistance for code-switching speech recognition. In *Proc. Interspeech 2022*, pages 3924–3928.

Sanket Sonu, Rejwanul Haque, Mohammed Hasanuzzaman, Paul Stynes, and Pramod Pathak. 2022. Identifying emotions in code mixed hindi-english tweets. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 35–41.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, pos tags and entrainment in code-switching. In *Proc. Interspeech 2018*, pages 1938–1942.

Victor Soto and Julia Hirschberg. 2017. Crowdsourcing universal part-of-speech tags for code-switching. In *Proc. Interspeech 2017*, pages 77–81.

Victor Soto and Julia Hirschberg. 2018. Joint part-of-speech and language id tagging for code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–10.

Victor Soto and Julia Hirschberg. 2019. Improving code-switched language modeling performance using cognate features. In *Proc. Interspeech 2019*, pages 3725–3729.

Mithun Kumar SR, Lov Kumar, and Aruna Malapati. 2022. Sentiment analysis on code-switched dravidian languages with kernel based extreme learning machines. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 184–190.

Dama Sravani, Lalitha Kameswari, and Radhika Mamidi. 2021. Political discourse analysis: a case study of code mixing and code switching in political speeches. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–5.

Anirudh Srinivasan. 2020. Msr india at semeval-2020 task 9: Multilingual models can do code-mixing too. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 951–956.

Anirudh Srinivasan, Sandipan Dandapat, and Monojit Choudhury. 2020. Code-mixed parse trees and how to find them. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 57–64.

Vamshi Krishna Srirangam, Appidi Abhinav Reddy, Vinay Singh, and Manish Shrivastava. 2019. Corpus creation and analysis for named entity recognition in telugu-english code-mixed social media data. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 183–189.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Aditya Srivastava and V Harsha Vardhan. 2020. Hcms at semeval-2020 task 9: A neural approach to sentiment analysis for code-mixed texts. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1253–1258.

Brij Mohan Lal Srivastava and Sunayana Sitaram. 2018. Homophone identification and merging for code-switched speech recognition. In *Proc. Interspeech 2018*, pages 1943–1947.

Vivek Srivastava and Mayank Singh. 2020a. Iit gandhinagar at semeval-2020 task 9: Code-mixed sentiment classification using candidate sentence generation and selection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1259–1264.

Vivek Srivastava and Mayank Singh. 2020b. Phinc: A parallel hinglish social media code-mixed corpus for machine translation. In *Proceedings of the Sixth*

Workshop on Noisy User-generated Text (W-NUT 2020), pages 41–49.

Vivek Srivastava and Mayank Singh. 2021a. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14.

Vivek Srivastava and Mayank Singh. 2021b. Hinge: A dataset for generation and evaluation of code-mixed hinglish text. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 200–208.

Vivek Srivastava and Mayank Singh. 2021c. Quality evaluation of the low-resource synthetically generated code-mixed hinglish text. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 314–319.

Sara Stymne et al. 2020. Evaluating word embeddings for indonesian–english code-mixed text based on synthetic data. In *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*, pages 26–35.

Ahmed Sultan, Mahmoud Salim, Amina Gaber, and Islam El Hosary. 2020. Wessa at semeval-2020 task 9: Code-mixed sentiment analysis using transformers. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1342–1347.

Charles Sutton, Andrew McCallum, et al. 2012. An introduction to conditional random fields. *Foundations and Trends® in Machine Learning*, 4(4):267–373.

Krithika Swaminathan, K Divyasri, GL Gayathri, Thenmozhi Durairaj, and B Bharathi. 2022. Pandas@ abusive comment detection in tamil code-mixed data using custom embeddings with labse. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119.

Chihiro Taguchi, Sei Iwata, and Taro Watanabe. 2022. Universal dependencies treebank for tatar: Incorporating intra-word code-switching information. In *Proceedings of the Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia within the 13th Language Resources and Evaluation Conference*, pages 95–104.

Chihiro Taguchi, Yusuke Sakai, and Taro Watanabe. 2021. Transliteration for low-resource code-switching texts: Building an automatic cyrillic-to-latin converter for tatar. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 133–140.

Karan Taneja, Satarupa Guha, Preethi Jyothi, and Basil Abraham. 2019. Exploiting monolingual speech corpora for code-mixed speech recognition. In *Proc. Interspeech 2019*, pages 2150–2154.

Ishan Tarunesh, Syamantak Kumar, and Preethi Jyothi. 2021. From machine translation to code-switching: Generating high-quality code-switched text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3154–3169.

Anju Leela Thomas, Anusha Prakash, Arun Baby, and Hema Murthy. 2018a. Code-switching in indic speech synthesisers. In *Proc. Interspeech 2018*, pages 1948–1952.

Anju Leela Thomas, Anusha Prakash, Arun Baby, and Hema A Murthy. 2018b. Code-switching in indic speech synthesisers. In *INTERSPEECH*, pages 1948–1952.

Jinchuan Tian, Jianwei Yu, Chunlei Zhang, Yuexian Zou, and Dong Yu. 2022. Lae: Language-aware encoder for monolingual and multilingual asr. In *Proc. Interspeech 2022*, pages 3178–3182.

Shashwat Trivedi, Harsh Rangwani, and Anil Kumar Singh. 2018. Iit (bhu) submission for the acl shared task on named entity recognition on code-switched data. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 148–153.

G Richard Tucker. 2001. A global perspective on bilingualism and bilingual education. *GEORGETOWN UNIVERSITY ROUND TABLE ON LANGUAGES AND LINGUISTICS 1999*, page 332.

Ewald van der Westhuizen and Thomas Niesler. 2017. Synthesising isizulu-english code-switch bigrams using word embeddings. In *INTERSPEECH*, pages 72–76.

Charangan Vasantharajan and Uthayasanker Thayasivam. 2021. Hypers@ dravidianlangtech-eacl2021: Offensive language identification in dravidian code-mixed youtube comments and posts. In *Proceedings of the first workshop on speech and language technologies for dravidian languages*, pages 195–202.

Deepanshu Vijay, Aditya Bohra, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. Corpus creation and emotion prediction for hindi-english code-mixed social media text. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: student research workshop*, pages 128–135.

David Vilares, Miguel A Alonso, and Carlos Gómez-Rodríguez. 2016. En-es-cs: An english-spanish code-switching twitter corpus for multilingual sentiment analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4149–4153.

Martin Volk and Simon Clematide. 2014. Detecting code-switching in a multilingual alpine heritage corpus. In *Proceedings of the first workshop on computational approaches to code switching*, pages 24–33.

Martin Volk, Lukas Fischer, Patricia Scheurer, Bernard Silvan Schroffenegger, Raphael Schwitter, Phillip Ströbel, and Benjamin Suter. 2022. Nunc profana tractemus. detecting code-switching in a large corpus of 16th century letters. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2901–2908.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014. Finding romanized arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2249–2253.

Ngoc Thang Vu and Tanja Schultz. 2014. Exploration of the impact of maximum entropy in recurrent neural network language models for code-switching speech. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 34–41.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 974–979.

Anshul Wadhawan and Akshita Aggarwal. 2021. Towards emotion recognition in hindi-english code-mixed data: A transformer based approach. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 195–202.

Changhan Wang, Kyunghyun Cho, and Douwe Kiela. 2018. Code-switched named entity recognition with embedding attention. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 154–158.

Jisung Wang, Jihwan Kim, Sangki Kim, and Yeha Lee. 2020. Exploring lexicon-free modeling units for end-to-end korean and korean-english code-switching speech recognition. In *Proc. Interspeech 2020*, pages 1072–1075.

Qinyi Wang, Emre Yılmaz, Adem Derinel, and Haizhou Li. 2019. Code-switching detection using asr-generated language posteriors. In *Proc. Interspeech 2019*, pages 3740–3744.

Zhongqing Wang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2015. Emotion detection in code-switching texts via bilingual and sentimental information. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 763–768.

Zhongqing Wang, Yue Zhang, Sophia Lee, Shoushan Li, and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1634.

Jochen Weiner, Ngoc Thang Vu, Dominic Telaar, Florian Metze, Tanja Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. 2012a. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Spoken Language Technologies for Under-Resourced Languages*.

Jochen Weiner, Ngoc Thang Vu, Dominic Telaar, Florian Metze, Tanja Schultz, Dau-Cheng Lyu, Eng-Siong Chng, and Haizhou Li. 2012b. Integration of language identification into a recognition system for spoken conversations containing code-switches. In *Proc. 3rd Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2012)*, pages 76–79.

Christopher M White, Sanjeev Khudanpur, and James K Baker. 2008. An investigation of acoustic models for multilingual code-switching. In *Ninth Annual Conference of the International Speech Communication Association*.

Matthew Wiesner, Mousmita Sarma, Ashish Arora, Desh Raj, Dongji Gao, Ruizhe Huang, Supreet Preet, Moris Johnson, Zikra Iqbal, Nagendra Goel, Jan Trmal, Leibny Paola García Perera, and Sanjeev Khudanpur. 2021. Training hybrid models on noisy transliterated transcripts for code-switched speech recognition. In *Proc. Interspeech 2021*, pages 2906–2910.

Nick Wilkinson, Astik Biswas, Emre Yilmaz, Febe De Wet, Thomas Niesler, et al. 2020. Semi-supervised acoustic modelling for five-lingual code-switched asr using automatically-segmented soap opera speech. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 70–78.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preoţiuc-Pietro. 2022. Cross-lingual few-shot learning on unseen languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 777–791.

Genta Indra Winata, Samuel Cahyawijaya, Zhaojiang Lin, Zihan Liu, Peng Xu, and Pascale Fung. 2020. Meta-transfer learning for code-switched speech recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3770–3776.

Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. 2021a. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153.

Genta Indra Winata, Zhaojiang Lin, and Pascale Fung. 2019a. Learning multilingual meta-embeddings for

code-switching named entity recognition. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 181–186.

Genta Indra Winata, Zhaojiang Lin, Jamin Shin, Zihan Liu, and Pascale Fung. 2019b. Hierarchical meta-embeddings for code-switching named entity recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3541–3547.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021b. Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018a. Code-switching language modeling using syntax-aware multi-task learning. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 62–67.

Genta Indra Winata, Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2019c. Code-switched language models using neural based synthetic data from parallel sentences. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 271–280.

Genta Indra Winata, Chien-Sheng Wu, Andrea Madotto, and Pascale Fung. 2018b. Bilingual character representation for efficiently addressing out-of-vocabulary words in code-switching named entity recognition. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 110–114.

Jane Wottawa, Amazouz Djegdjiga, Martine Adda-Decker, and Lori Lamel. 2018. Studying vowel variation in french-algerian arabic code-switched speech. In *Proc. Interspeech 2018*, pages 2753–2757.

Qi Wu, Peng Wang, and Chenghao Huang. 2020. Meistermorxrc at semeval-2020 task 9: Fine-tune bert and multitask learning for sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1294–1297.

Yi-Lun Wu, Chaio-Wen Hsieh, Wei-Hsuan Lin, Chun-Yi Liu, and Liang-Chih Yu. 2011. Unknown word extraction from multilingual code-switching sentences. In *ROCLING 2011 Poster Papers*, pages 349–360, Taipei, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Meng Xuan Xia. 2016. Codeswitching language identification using subword information enriched word vectors. In *Proceedings of the second workshop on computational approaches to code switching*, pages 132–136.

Meng Xuan Xia and Jackie Chi Kit Cheung. 2016. Accurate pinyin-english codeswitched language identification. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 71–79.

Xinyuan Xia, Lu Xiao, Kun Yang, and Yueyue Wang. 2022. Identifying tension in holocaust survivors' interview: Code-switching/code-mixing as cues. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1490–1495.

Haihua Xu, Van Tung Pham, Zin Tun Kyaw, Zhi Hao Lim, Eng Siong Chng, and Haizhou Li. 2018. Mandarin-english code-switching speech recognition. In *Proc. Interspeech 2018*, pages 554–555.

Jitao Xu and François Yvon. 2021. Can you traducir this? machine translation for code-switched input. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 84–94.

Qihui Xu, Magdalena Markowska, Martin Chodorow, and Ping Li. 2021. A network science approach to bilingual code-switching. *Proceedings of the Society for Computation in Linguistics*, 4(1):18–27.

Liumeng Xue, Wei Song, Guanghui Xu, Lei Xie, and Zhizheng Wu. 2019. Building a mixed-lingual neural tts system with only monolingual data. In *Proc. Interspeech 2019*, pages 2060–2064.

Zhen Yang, Bojie Hu, Ambyera Han, Shen Huang, and Qi Ju. 2020. Csp: Code-switching pre-training for neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2624–2636.

Lingxuan Ye, Gaofeng Cheng, Runyan Yang, Zehui Yang, Sanli Tian, Pengyuan Zhang, and Yonghong Yan. 2022. Improving recognition of out-of-vocabulary words in e2e code-switching asr by fusing speech generation methods. In *Proc. Interspeech 2022*, pages 3163–3167.

Yin-Lai Yeong and Tien-Ping Tan. 2010. Language identification of code switching malay-english words using syllable structure information. In *Proc. Spoken Language Technologies for Under-Resourced Languages*, pages 142–145.

Yin-Lai Yeong and Tien-Ping Tan. 2014. Language identification of code switching sentences and multilingual sentences of under-resourced languages by using multi structural word information. In *Proc. Interspeech 2014*, pages 3052–3055.

E Yilmaz, H Heuvel, and DA van Leeuwen. 2018. Acoustic and textual data augmentation for improved asr of code-switching speech. In *Proceedings of Interspeech*, pages 1933–1937. Hyderabad, India: ISCA.

Emre Yilmaz, Maaike Andringa, Sigrid Kingma, Jelske Dijkstra, Frits van der Kuip, Hans Van de Velde, Frederik Kampstra, Jouke Algra, Henk van den Heuvel, and David van Leeuwen. 2016. A longitudinal bilingual frisian-dutch radio broadcast database designed for code-switching research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4666–4669.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a unified code-switching asr system for south african languages. *Proceedings of Interspeech*.

Emre Yilmaz, Henk Van Den Heuvel, and David Van Leeuwen. 2018. Code-switching detection with data-augmented acoustic and language models. In *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 127–131.

Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between turkish-english language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Genta Indra Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, Yin Lin Tan, and Alham Fikri Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages.

Shan-Ruei You, Shih-Chieh Chien, Chih-Hsing Hsu, Ke-Shiu Chen, Jia-Jang Tu, Jeng Shien Lin, and Sen-Chia Chang. 2004. Chinese-english mixed-lingual keyword spotting. In *2004 International Symposium on Chinese Spoken Language Processing*, pages 237–240. IEEE.

Fu-Hao Yu and Kuan-Yu Chen. 2020. A preliminary study on leveraging meta learning technique for code-switching speech recognition. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 136–147.

Liang-Chih Yu, Wei-Cheng He, and Wei-Nan Chien. 2012. A language modeling approach to identifying code-switched sentences and words. In *Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 3–8.

Emre Yılmaz, Samuel Cohen, Xianghu Yue, David A. van Leeuwen, and Haizhou Li. 2019. Multi-graph decoding for code-switching asr. In *Proc. Interspeech 2019*, pages 3750–3754.

Emre Yılmaz, Jelske Dijkstra, Hans Van de Velde, Frederik Kampstra, Jouke Algra, Henk van den Heuvel, and David Van Leeuwen. 2017a. Longitudinal speaker clustering and verification corpus with code-switching frisian-dutch speech. In *Proc. Interspeech 2017*, pages 37–41.

Emre Yılmaz, Henk van den Heuvel, Jelske Dijkstra, Hans Van de Velde, Frederik Kampstra, Jouke Algra, and David Van Leeuwen. 2016. Open source speech and language resources for frisian. In *Proc. Interspeech 2016*, pages 1536–1540.

Emre Yılmaz, Henk van den Heuvel, and David Van Leeuwen. 2017b. Exploiting untranscribed broadcast data for improved code-switching detection. In *Proc. Interspeech 2017*, pages 42–46.

Emre Yılmaz, Henk van den Heuvel, and David van Leeuwen. 2018. Acoustic and textual data augmentation for improved asr of code-switching speech. In *Proc. Interspeech 2018*, pages 1933–1937.

George-Eduard Zaharia, George-Alexandru Vlad, Dumitru-Clementin Cercel, Traian Rebedea, and Costin Chiru. 2020. Upb at semeval-2020 task 9: Identifying sentiment in code-mixed social media texts using transformers and multi-task learning. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1322–1330.

Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, and Haizhou Li. 2019. On the end-to-end solution to mandarin-english code-switching speech recognition. In *Proc. Interspeech 2019*, pages 2165–2169.

Haobo Zhang, Haihua Xu, Van Tung Pham, Hao Huang, and Eng Siong Chng. 2020. Monolingual data selection analysis for english-mandarin hybrid code-switching speech recognition. In *Proc. Interspeech 2020*, pages 2392–2396.

Shiliang Zhang, Yuan Liu, Ming Lei, Bin Ma, and Lei Xie. 2019. Towards language-universal mandarin-english speech recognition. In *Proc. Interspeech 2019*, pages 2170–2174.

Shuai Zhang, Jiangyan Yi, Zhengkun Tian, Ye Bai, Jianhua Tao, Xuefei Liu, and Zhengqi Wen. 2021a. End-to-end spelling correction conditioned on acoustic feature for code-switching speech recognition. In *Proc. Interspeech 2021*, pages 266–270.

Shuai Zhang, Jiangyan Yi, Zhengkun Tian, Jianhua Tao, Yu Ting Yeung, and Liqun Deng. 2022. reducing multilingual context confusion for end-to-end code-switching automatic speech recognition. In *Proc. Interspeech 2022*, pages 3894–3898.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021b. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9220–9230.

Yi Zhang and Jian-Hua Tao. 2008. Prosody modification on mixed-language speech synthesis. In *Proc. International Symposium on Chinese Spoken Language Processing*, pages 253–256.

Shengkui Zhao, Trung Hieu Nguyen, Hao Wang, and Bin Ma. 2020. Towards natural bilingual and code-switched speech synthesis based on mix of monolingual recordings and cross-lingual voice conversion. In *Proc. Interspeech 2020*, pages 2927–2931.

Xinyuan Zhou, Emre Yılmaz, Yanhua Long, Yijie Li, and Haizhou Li. 2020. Multi-encoder-decoder transformer for code-switching speech recognition. In *Proc. Interspeech 2020*, pages 1042–1046.

Yueying Zhu, Xiaobing Zhou, Hongling Li, and Kunjie Dong. 2020. Zyy1510 team at semeval-2020 task 9: Sentiment analysis for code-mixed social media text with sub-word level representations. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1354–1359.

A Annotation Catalog

We release the annotation of all papers we use in the survey.

A.1 *CL Anthology

Bilingual Table 7 shows the annotation for papers with African-English. Table 8 shows the annotation for papers with East Asian-English languages. Table 9 shows the annotation for papers with European-English languages. Table 10 shows the annotation for papers with Middle Eastern-English languages. Table 11 and Table 12 show the annotation for papers with South Asian-English languages. Table 13 shows the annotation for papers with South East Asian-English languages. Table 14 shows the annotation for papers with a combination of a language with a dialect. Table 15 shows the annotation for papers with languages in the same family. Table 16 shows the annotation for papers with languages in different families.

Trilingual Table 17 shows the annotation for papers with three languages.

4+ Table 18 shows the annotation for papers with four or more languages.

A.2 ISCA Proceeding

Bilingual Table 19 shows the annotation for papers with African-English. Table 20 shows the annotation for papers with East Asian-English languages. Table 21 shows the annotation for papers with European-English languages. Table 22 shows the annotation for papers with Middle Eastern-English languages. Table 23 shows the annotation for papers with South Asian-English languages. Table 24 shows the annotation for papers with South East Asian-English languages. Table 25 shows the annotation for papers with a combination of a language with a dialect. Table 26 shows the annotation for papers with languages in the same family. Table 27 shows the annotation for papers with languages in different families.

Trilingual Table 28 shows the annotation for papers with three languages.

4+ Table 29 shows the annotation for papers with four or more languages.

Paper	Proceeding	IsiZulu	Swahili	isiXhosa	Setswana	Sesotho
		5	1	3	3	3
(Joshi, 1982)	COLING	✓				
(Piergallini et al., 2016)	CALCS		✓			
(Niesler et al., 2018)	LREC	✓		✓	✓	✓
(Biswas et al., 2020)	CALCS	✓				
(Wilkinson et al., 2020)	SLTU and CCURL	✓		✓	✓	✓
(Biswas et al., 2020)	LREC	✓		✓	✓	✓

Table 7: *CL Catalog in African-English.

Paper	Proceeding	Chinese	Cantonese	Korean
		20	1	1
(Fung et al., 1999)	ACL	✓		
(Chan et al., 2009)	IJCLCLP		✓	
(Li et al., 2012)	LREC	✓		
(Peng et al., 2014)	ACL-IJCNLP	✓		
(Li and Fung, 2014)	EMNLP	✓		
(Solorio et al., 2014)	CALCS	✓		
(Chittaranjan et al., 2014)	CALCS	✓		
(Lin et al., 2014)	CALCS	✓		
(Jain and Bhat, 2014)	CALCS	✓		
(King et al., 2014)	CALCS	✓		
(Huang and Yates, 2014)	EACL	✓		
(Wang et al., 2015)	ACL-IJCNLP	✓		
(Gambäck and Das, 2016)	LREC	✓		
(Wang et al., 2016)	COLING	✓		
(Çetinoğlu et al., 2016)	CALCS	✓		
(Xia and Cheung, 2016)	CALCS	✓		
(Yang et al., 2020)	EMNLP	✓		
(Calvillo et al., 2020)	EMNLP	✓		
(Lin and Chen, 2020)	ROCLING	✓		
(Cho et al., 2020)	CALCS			✓
(Lin and Chen, 2021)	ROCLING	✓		
(Lovenia et al., 2021)	LREC	✓		

Table 8: *CL Catalog in East Asian-English.

Paper	Proceeding	Spanish	French	Portugese	Polish	German	Dutch	Finnish
		78	7	1	1	5	2	1
(Sankoff, 1998)	COLING							✓
(Solorio and Liu, 2008a)	EMNLP	✓						
(Solorio and Liu, 2008b)	EMNLP	✓						
(Peng et al., 2014)	ACL-IJCNLP	✓						
(Solorio et al., 2014)	CALCS	✓						
(Chittaranjan et al., 2014)	CALCS	✓						
(Lin et al., 2014)	CALCS	✓						
(Jain and Bhat, 2014)	CALCS	✓						
(King et al., 2014)	CALCS	✓						
(Carpuat, 2014)	CALCS		✓					
(Barman et al., 2014b)	CALCS	✓						
(Shrestha, 2014)	CALCS	✓						
(Bar and Dershowitz, 2014)	CALCS	✓						
(Gambäck and Das, 2016)	LREC	✓						
(Vilares et al., 2016)	LREC	✓						
(Çetinoğlu et al., 2016)	CALCS	✓						
(Guzman et al., 2016)	CALCS	✓						
(Molina et al., 2016)	CALCS	✓						
(Samih et al., 2016a)	CALCS	✓						
(Jaech et al., 2016)	CALCS	✓						
(AlGhamdi et al., 2016)	CALCS	✓						
(Al-Badrashiny and Diab, 2016)	CALCS	✓						
(Chanda et al., 2016a)	CALCS	✓						
(Shirvani et al., 2016)	CALCS	✓						
(Shrestha, 2016)	CALCS	✓						
(Sikdar and Gambäck, 2016)	CALCS	✓						
(Xia, 2016)	CALCS	✓						
(Duong et al., 2017)	CoNLL					✓		
(Rijhwani et al., 2017)	ACL	✓	✓	✓			✓	
(Choudhury et al., 2017)	ICON	✓						
(Núñez and Wisniewski, 2018)	TALN PFIA	✓						
(Pratapa et al., 2018b)	EMNLP	✓						
(Mendels et al., 2018)	LREC	✓						
(Soto and Hirschberg, 2018)	CALCS	✓						
(Mave et al., 2018)	CALCS	✓						
(Bullock et al., 2018a)	CALCS	✓						
(Rallabandi et al., 2018)	CALCS	✓						
(Bawa et al., 2018)	CALCS	✓						
(Jain et al., 2018)	CALCS	✓						
(Winata et al., 2018b)	CALCS	✓						
(Sikdar et al., 2018)	CALCS	✓						
(Janke et al., 2018)	CALCS	✓						
(Geetha et al., 2018)	CALCS	✓						
(Claeser et al., 2018)	CALCS	✓						
(Aguilar et al., 2018)	CALCS	✓						
(Trivedi et al., 2018)	CALCS	✓						
(Wang et al., 2018)	CALCS	✓						
(Gonen and Goldberg, 2019)	EMNLP	✓						
(Yang et al., 2020)	EMNLP		✓			✓		
(Khanuja et al., 2020b)	ACL	✓						
(Aguilar and Solorio, 2020)	ACL	✓						
(Cameron, 2020)	JEP		✓					
(Ahn et al., 2020)	SCiL	✓						
(Srinivasan et al., 2020)	CALCS	✓						
(Patwa et al., 2020)	SemEval	✓						
(De Leon et al., 2020)	SemEval	✓						
(Aparaschivei et al., 2020)	SemEval	✓						
(Kong et al., 2020)	SemEval	✓						
(Angel et al., 2020)	SemEval	✓						
(Palomino and Ochoa-Luna, 2020)	SemEval	✓						
(Ma et al., 2020)	SemEval	✓						
(Kumar et al., 2020)	SemEval	✓						
(Advani et al., 2020)	SemEval	✓						
(Javdan et al., 2020)	SemEval	✓						
(Wu et al., 2020)	SemEval	✓						
(Zaharia et al., 2020)	SemEval	✓						
(Sultan et al., 2020)	SemEval	✓						
(Zhu et al., 2020)	SemEval	✓						
(Parekh et al., 2020)	CoNLL	✓						
(Gupta et al., 2020)	Findings of EMNLP	✓	✓			✓		
(Aguilar et al., 2020)	LREC	✓						
(Iliescu et al., 2021)	CALCS	✓						
(Xu and Yvon, 2021)	CALCS	✓	✓					
(Gupta et al., 2021b)	CALCS	✓						
(Jayanthi et al., 2021)	CALCS	✓						
(Winata et al., 2021a)	CALCS	✓						
(Prasad et al., 2021)	MRL	✓						
(Chopra et al., 2021)	Findings of EMNLP	✓						
(Santy et al., 2021)	AdaptNLP	✓						
(Cheong et al., 2021)	W-NUT	✓						
(Pratapa and Choudhury, 2021)	W-NUT	✓	✓			✓	✓	
(Xia et al., 2022)	LREC				✓	✓		
(Mellado and Lignos, 2022)	LREC	✓						
(Ostapenko et al., 2022)	ACL	✓						

Table 9: *CL Catalog in European-English.

Paper	Proceeding	Egyptian Arabic	Arabic	Turkish
		3	1	2
(Rijhwani et al., 2017)	ACL			✓
(Hamed et al., 2018)	LREC	✓		
(Yirmibeşoğlu and Eryiğit, 2018)	W-NUT			✓
(Sabty et al., 2020)	WANLP		✓	
(Balabel et al., 2020)	LREC	✓		
(Hamed et al., 2020)	LREC	✓		

Table 10: *CL Catalog in Middle Eastern-English.

Paper	Proceeding	Hindi	Marathi	Konkani	Bengali	Bengali intra-word	Nepali	Telugu	Bangla	Gujarati	Punjabi	Tamil	Malayalam	Malayalam scripts	Kannada
		111	1	1	12	1	10	7	1	1	2	37	23	1	10
(Joshi, 1982)	COLING		✓												
(Sankoff, 1998)	COLING											✓			
(Bhattacharja, 2010)	PACLIC					✓									
(Diab and Kamboj, 2011)	ALR	✓													
(Dey and Fung, 2014)	LREC	✓													
(Das and Gambäck, 2014)	ICON	✓			✓										
(Vyas et al., 2014)	EMNLP	✓													
(Jhamtani et al., 2014)	PACLIC	✓													
(Barman et al., 2014a)	CALCS	✓			✓										
(Solorio et al., 2014)	CALCS						✓								
(Chittaranjan et al., 2014)	CALCS						✓								
(Lin et al., 2014)	CALCS						✓								
(Jain and Bhat, 2014)	CALCS						✓								
(King et al., 2014)	CALCS						✓								
(Bali et al., 2014)	CALCS	✓													
(Barman et al., 2014b)	CALCS						✓								
(Shrestha, 2014)	CALCS						✓								
(Jamatia et al., 2015)	RANLP	✓													
(Sequiera et al., 2015)	ICON	✓													
(Gupta et al., 2016)	ICON	✓													
(Asnani and Pawar, 2016)	ICON	✓													
(Begum et al., 2016)	LREC	✓													
(Gambäck and Das, 2016)	LREC	✓					✓								
(Sitaram and Black, 2016)	LREC	✓													
(Sharma et al., 2016)	HLT-NAACL	✓													
(Joshi et al., 2016)	COLING	✓													
(Çetinoğlu et al., 2016)	CALCS	✓													
(Chanda et al., 2016b)	CALCS				✓										
(Ghosh et al., 2016)	CALCS	✓			✓							✓			
(Pratapa and Choudhury, 2017)	ICON	✓													
(Phadte and Thakkar, 2017)	ICON			✓											
(Pandey et al., 2017)	ICON	✓													
(Bhat et al., 2017)	EACL	✓													
(Banerjee et al., 2018)	COLING	✓			✓					✓		✓			
(Gundapu and Mamidi, 2018)	PACLIC							✓							
(Ball and Garrette, 2018)	EMNLP	✓													
(Khandelwal et al., 2018)	LREC	✓													
(Kumar et al., 2018)	LREC	✓													
(Pandey et al., 2018)	LREC	✓													
(Hartmann et al., 2018)	LREC	✓													
(Gupta et al., 2018a)	LREC	✓										✓			
(Mandal and Nanmaran, 2018)	W-NUT				✓										
(Mandal and Singh, 2018)	W-NUT	✓			✓										
(Vijay et al., 2018)	SRW	✓													
(Gupta et al., 2018b)	CoNLL	✓													
(Singh et al., 2018)	SRW	✓													
(Chandu et al., 2019)	CALCS	✓						✓				✓			
(Sivasankaran et al., 2018)	CALCS	✓													
(Chandu et al., 2019)	CALCS	✓						✓				✓			
(Gupta et al., 2018c)	CALCS	✓													
(Mave et al., 2018)	CALCS	✓													
(Rallabandi et al., 2018)	CALCS	✓													
(Bawa et al., 2018)	CALCS	✓													
(Chandu et al., 2018)	CALCS	✓													
(Jain et al., 2018)	CALCS	✓													
(Mukherjee et al., 2019)	ICON	✓													
(Raha et al., 2019)	ICON				✓										
(Shah et al., 2019)	AnnoNLP	✓													
(Ghosh et al., 2019)	TLT, SyntaxFest	✓			✓										
(Srirangam et al., 2019)	SRW							✓							
(Lal et al., 2019)	SRW	✓													
(Chakravarthi, 2020)	PEOPLES											✓	✓		
(Singh and Lefever, 2020)	ICON										✓				
(Bansal et al., 2020a)	ICON										✓				
(Bansal et al., 2020c)	ACL	✓													
(Khanuja et al., 2020b)	ACL	✓													
(Aguilar and Solorio, 2020)	ACL	✓					✓								
(Rani et al., 2020)	TRAC	✓													
(Pant and Dadu, 2020)	SRW	✓													
(Khanuja et al., 2020a)	CALCS	✓													
(Singh and Lefever, 2020)	CALCS	✓													
(Srinivasan et al., 2020)	CALCS	✓													
(Patwa et al., 2020)	SemEval	✓													
(Liu et al., 2020)	SemEval	✓													
(Aparaschivei et al., 2020)	SemEval	✓													
(Kong et al., 2020)	SemEval	✓													
(Baruah et al., 2020)	SemEval	✓													
(Srinivasan, 2020)	SemEval	✓													
(Goswami et al., 2020)	SemEval	✓													
(Kumar et al., 2020)	SemEval	✓													
(Advani et al., 2020)	SemEval	✓													
(dos Santos Neto et al., 2020)	SemEval	✓													
(Gundapu and Mamidi, 2020)	SemEval	✓													
(Srivastava and Vardhan, 2020)	SemEval	✓													
(Srivastava and Singh, 2020a)	SemEval	✓													
(Parikh et al., 2020)	SemEval	✓													
(Javdan et al., 2020)	SemEval	✓													
(Garain et al., 2020)	SemEval	✓													
(Banerjee et al., 2020)	SemEval	✓													

Table 11: *CL Catalog in South Asian-English (1).

(Wu et al., 2020)	SemEval	✓						
(Baroi et al., 2020)	SemEval	✓						
(Gopalan and Hopkins, 2020)	SemEval	✓						
(Malte et al., 2020)	SemEval	✓						
(Zaharia et al., 2020)	SemEval	✓						
(Zhu et al., 2020)	SemEval	✓						
(Parekh et al., 2020)	CoNLL	✓						
(Chakravarthi et al., 2020a)	SLTU & CCURL						✓	
(Chakravarthi et al., 2020b)	SLTU & CCURL					✓		
(Gupta et al., 2020)	Findings of EMNLP	✓	✓		✓	✓	✓	
(Makhija et al., 2020)	COLING	✓						
(Aguilar et al., 2020)	LREC	✓		✓				
(Chatterjere et al., 2020)	LREC	✓						
(Aggarwal et al., 2020)	W-NUT	✓						
(Srivastava and Singh, 2020b)	W-NUT	✓						
(Chakravarthy et al., 2020)	W-NUT	✓						
(Chinnappa, 2021)	LTEDI					✓	✓	
(Dave et al., 2021)	LTEDI					✓	✓	
(Hossain et al., 2021)	LTEDI					✓	✓	
(Balouchzahi et al., 2021)	LTEDI					✓	✓	
(Agarwal and Narula, 2021)	SRW	✓						
(Agarwal et al., 2021)	NLP4ConvAI	✓						
(Garg et al., 2021)	Eval4NLP	✓						
(Srivastava and Singh, 2021b)	Eval4NLP	✓						
(Tarunesh et al., 2021)	ACL	✓						
(Srivastava and Singh, 2021a)	CALCS	✓						
(Gautam et al., 2021a)	CALCS	✓						
(Dowlagar and Mamidi, 2021a)	CALCS	✓						
(Appicharla et al., 2021)	CALCS	✓						
(Jawahar et al., 2021)	CALCS	✓						
(Gautam et al., 2021b)	CALCS	✓						
(Gupta et al., 2021b)	CALCS	✓				✓	✓	
(Jayanthi et al., 2021)	CALCS	✓				✓		
(Parikh and Solorio, 2021)	CALCS	✓						
(Winata et al., 2021a)	CALCS	✓						
(Mehnaz et al., 2021)	EMNLP	✓						
(Prasad et al., 2021)	MRL	✓				✓	✓	
(Gupta et al., 2021a)	NAACL	✓						
(Ravikiran and Annamalai, 2021)	DravidianLangTech					✓		✓
(Mahata et al., 2021)	DravidianLangTech					✓		
(Saumya et al., 2021)	DravidianLangTech					✓	✓	✓
(Mandalam and Sharma, 2021)	DravidianLangTech					✓	✓	
(Dowlagar and Mamidi, 2021b)	DravidianLangTech					✓	✓	
(Gupta et al., 2021c)	DravidianLangTech					✓	✓	
(Balouchzahi and Shashirekha, 2021)	DravidianLangTech					✓	✓	
(Dowlagar and Mamidi, 2021c)	DravidianLangTech					✓	✓	✓
(Li, 2021)	DravidianLangTech					✓	✓	✓
(Andrew, 2021)	DravidianLangTech					✓	✓	✓
(Vasantharajan and Thayasivam, 2021)	DravidianLangTech					✓	✓	✓
(Huang and Bai, 2021)	DravidianLangTech					✓	✓	✓
(Sharif et al., 2021)	DravidianLangTech					✓	✓	✓
(Bharathi et al., 2021)	DravidianLangTech					✓	✓	✓
(Balouchzahi et al., 2021)	DravidianLangTech					✓	✓	✓
(Rajalakshmi et al., 2021)	DravidianLangTech					✓		
(Khan et al., 2021)	Findings of EMNLP	✓						
(Chopra et al., 2021)	Findings of EMNLP	✓	✓		✓			
(Srivastava and Singh, 2021c)	INLG	✓						
(Santy et al., 2021)	AdaptNLP	✓						
(Wadhawan and Aggarwal, 2021)	WASSA	✓						
(Priyanshu et al., 2021)	W-NUT	✓						
(Dutta, 2022)	DCLRL				✓			
(Biradar and Saumya, 2022)	DravidianLangTech					✓		
(Swaminathan et al., 2022)	DravidianLangTech					✓		
(SR et al., 2022)	DravidianLangTech					✓	✓	✓
(Ravikiran and Chakravarthi, 2022)	DravidianLangTech					✓		
(Ravikiran et al., 2022)	DravidianLangTech					✓		
(Nayak and Joshi, 2022)	WILDRE-6	✓						
(Gautam, 2022)	WILDRE-6	✓	✓		✓			
(Sonu et al., 2022)	WILDRE-6	✓						

Table 12: *CL Catalog in South Asian-English (2).

Paper	Proceeding	Vietnamese	Tagalog	Indonesian
		1	2	2
(Oco and Roxas, 2012)	PACLIC		✓	
(Stymne et al., 2020)	CALCS			✓
(Nguyen and Bryant, 2020)	LREC	✓		
(Arianto and Budi, 2020)	PACLIC			✓
(Herrera et al., 2022)	LREC			✓

Table 13: *CL Catalog in South East Asian-English.

Paper	Proceeding	Darija-MSA	MSA-Egyptian	MSA-Other Dialect	Chinese-Taiwanese	MSA-Levant Arabic	MSA-Gulf	Mixed-English
		1	15	10	2	2	1	1
(Chu et al., 2007)				✓				
(Yu et al., 2012)	CIPS-SIGHAN				✓			
(Elfardy and Diab, 2012)	COLING		✓			✓		
(Solorio et al., 2014)	CALCS			✓				
(Chittaranjan et al., 2014)	CALCS			✓				
(Lin et al., 2014)	CALCS			✓				
(Jain and Bhat, 2014)	CALCS			✓				
(Elfardy et al., 2014)	CALCS			✓				
(King et al., 2014)	CALCS			✓				
(Gambäck and Das, 2016)	LREC		✓					
(Samih and Maier, 2016)	LREC	✓						
(Diab et al., 2016)	LREC		✓					
(Molina et al., 2016)	CALCS		✓					
(Samih et al., 2016a)	CALCS		✓					
(Jaech et al., 2016)	CALCS			✓				
(Samih et al., 2016b)	CALCS			✓				
(AlGhamdi et al., 2016)	CALCS		✓					
(Al-Badrashiny and Diab, 2016)	CALCS			✓				
(Shrestha, 2016)	CALCS			✓				
(El-Haj et al., 2018)	LREC		✓			✓	✓	
(Shoemark et al., 2018)	W-NUT							✓
(Attia et al., 2018)	CALCS		✓					
(Janke et al., 2018)	CALCS		✓					
(Geetha et al., 2018)	CALCS		✓					
(Aguilar et al., 2018)	CALCS		✓					
(Wang et al., 2018)	CALCS		✓					
(Aguilar et al., 2020)	LREC		✓					
(Elmadany et al., 2021)	CALCS		✓					
(Winata et al., 2021a)	CALCS		✓					

Table 14: *CL Catalog in Dialect.

Paper	Proceeding	Komi-Zyrian - Russian	Arabizi-Arabic	Spanish-Catalan	Corsican-French	Frisian-Dutch
		1	1	1	1	3
(Eskander et al., 2014)	CALCS		✓			
(Yilmaz et al., 2016)	LREC					✓
(Braggaar and van der Goot, 2021)	AdaptNLP					✓
(Amin et al., 2022)	BioNLP			✓		
(Özateş et al., 2022)	Findings of NAACL	✓				✓
(Kevers, 2022)	SIGUL				✓	

Table 15: *CL Catalog in Two Languages in the same family.

Paper	Proceeding	Russian-Tatar	Russian-Tatar intra-word	Turkish-German	MSA-North African	French - Arabic Dialect	Dutch-Turkish	French-Algerian	Basque-Spanish	Spanish–Wixarika intra-word
		1	1	7	1	2	2	1	1	1
(Sankoff, 1998)	COLING					✓				
(Papalexakis et al., 2014)	CALCS						✓			
(Gambäck and Das, 2016)	LREC						✓			
(Çetinoğlu, 2016)	LREC			✓						
(Çetinoğlu et al., 2016)	CALCS			✓						
(Djegdjiga et al., 2018)	LREC							✓		
(El-Haj et al., 2018)	LREC				✓					
(Çetinoğlu and Çöltekin, 2019)	TLT, SyntaxFest 2019			✓						
(Mager et al., 2019)	NAACL			✓						✓
(Özateş and Çetinoğlu, 2021)	CALCS			✓						
(Taguchi et al., 2021)	CALCS	✓								
(Lounnas et al., 2021)	ICNLSP					✓				
(Aguirre et al., 2022)	LREC								✓	
(Özateş et al., 2022)	Findings of NAACL			✓						
(Taguchi et al., 2022)	EURALI		✓	✓						

Table 16: *CL Catalog in different family.

Paper	Proceeding	Tulu-Kannada-EN	Hindi-Bengali-EN	Greek-German-EN	Magahi-Hindi-EN	Arabic-EN-French	Darija-EN-French
		1	1	1	1	1	1
(Voss et al., 2014)	LREC						✓
(Çetinoğlu et al., 2016)	CALCS			✓			
(Barman et al., 2016)	CALCS		✓				
(Abdul-Mageed et al., 2020)	EMNLP					✓	
(Taguchi et al., 2021)	CALCS						
(Rani et al., 2022)	LREC				✓		
(Hegde et al., 2022)	ELRA	✓					

Table 17: *CL Catalog in Trilingual.

Paper	Proceeding	SEA Mandarin-English	Bangla-Chinese-Dutch-English-Farsi-German-Hindi-Korean-Russian-Spanish-Turkish	Early New High German, Latin, French, Greek, Italian, Hebrew	Telugu, Modern Standard Telugu, English, Hindi, Urdu,	MSA, Berber, French, local Algerian Arabic	Others (4+)	English, Swiss German Latin	Algerian, MSA, local Arabic varieties, Berber, French, and English	Mandarin-Hakka-Taiwanese-English
		10	4	1	1	2	3	1	1	1
(Wu et al., 2011)	ROCLING									✓
(Adel et al., 2013)	ACL	✓								
(Volk and Clematide, 2014)	CALCS							✓		
(Vu and Schultz, 2014)	CALCS	✓								
(Jurgens et al., 2014)	CALCS						✓			
(Garg et al., 2018a)	EMNLP	✓								
(Adouane et al., 2018)	CALCS						✓			
(Winata et al., 2018a)	CALCS	✓								
(Winata et al., 2019c)	CoNLL	✓								
(Lee and Li, 2020)	ACL	✓								
(Winata et al., 2020)	ACL	✓								
(Yu and Chen, 2020)	ROCLING	✓								
(Adouane and Bernardy, 2020)	CALCS					✓			✓	
(Adouane et al., 2020)	LREC					✓				
(Sravani et al., 2021)	CALCS				✓					
(Xu et al., 2021)	SCiL	✓								
(Zhang et al., 2021b)	EMNLP						✓			
(Cheong et al., 2021)	W-NUT	✓								
(Malmasi et al., 2022b)	SemEval		✓							
(Malmasi et al., 2022a)	COLING		✓							
(Volk et al., 2022)	LREC			✓						
(El Mekki et al., 2022)	SemEval		✓							
(Dowlagar and Mamidi, 2022)	SemEval		✓							

Table 18: *CL Catalog in 4+ languages.

Paper	Proceeding	isiZulu	isiXhosa	Setsawa	Sesotho	Sotho
		6	4	3	3	1
(Niesler and de Wet, 2008)	Odyssey	✓	✓			
(Mabokela et al., 2014)	SLTU					✓
(van der Westhuizen and Niesler, 2017)	Interspeech	✓				
(Yılmaz et al., 2018)	Interspeech	✓	✓	✓	✓	
(Biswas et al., 2018a)	Interspeech	✓				
(Biswas et al., 2018b)	SLTU	✓	✓	✓	✓	
(Biswas et al., 2019)	Interspeech	✓	✓	✓	✓	

Table 19: ISCA Catalog in African-English.

Paper	Proceeding	Chinese	Cantonese	Korean	Japanese
		27	5	1	1
(Fu and Shen, 2000)	ISCSLP	✓			
(Kojima and Tanaka, 2003)	Eurospeech				✓
(You et al., 2004)	ISCSLP	✓			
(Chan et al., 2004)	ISCSLP		✓		
(Chan et al., 2005)	Interspeech		✓		
(Ren et al., 2005)	Interspeech	✓			
(Chan et al., 2006)	Interspeech		✓		
(Liang et al., 2007)	SSW	✓			
(White et al., 2008)	Interspeech	✓			
(Qian et al., 2008)	ISCSLP	✓			
(Gu et al., 2008)	ISCSLP		✓		
(Zhang and Tao, 2008)	ISCSLP	✓			
(Cao et al., 2009)	Interspeech		✓		
(Shuang et al., 2010)	Interspeech	✓			
(He et al., 2012)	Interspeech	✓			
(Liang et al., 2013)	Interspeech	✓			
(Li and Fung, 2013)	Interspeech	✓			
(Xue et al., 2019)	Interspeech	✓			
(Gao et al., 2019)	Interspeech	✓			
(Zhang et al., 2019)	Interspeech	✓			
(Lu et al., 2020)	Interspeech	✓			
(Hu et al., 2020)	Interspeech	✓			
(Fu et al., 2020)	Interspeech	✓			
(Wang et al., 2020)	Interspeech			✓	
(Zhang et al., 2020)	Interspeech	✓			
(Chandu and Black, 2020)	Interspeech	✓			
(Zhao et al., 2020)	Interspeech	✓			
(Zhang et al., 2021a)	Interspeech	✓			
(Shen and Guo, 2022)	Interspeech	✓			
(Ye et al., 2022)	Interspeech	✓			
(Tian et al., 2022)	Interspeech	✓			
(Song et al., 2022)	Interspeech	✓			
(Zhang et al., 2022)	Interspeech	✓			
(Li et al., 2022)	Interspeech	✓			

Table 20: ISCA Catalog in East Asian-English.

Paper	Proceeding	Spanish	French	German	Maltese
(Pfister and Romsdorfer, 2003)	Eurospeech			✓	
(Romsdorfer and Pfister, 2005)	Interspeech		✓		
(Rosner and Farrugia, 2007)	Interspeech				✓
(Piccinini and Garellek, 2014)	SpeechProsody	✓			
(Sitaram et al., 2016)	SSW			✓	
(Soto and Hirschberg, 2017)	Interspeech	✓			
(Ramanarayanan and Suendermann-Oeft, 2017)	Interspeech	✓			
(Guzmán et al., 2017)	Interspeech	✓			
(Bullock et al., 2018b)	Interspeech	✓			
(Soto et al., 2018)	Interspeech	✓			
(Soto and Hirschberg, 2019)	Interspeech	✓			
(Chandu and Black, 2020)	Interspeech	✓			

Table 21: ISCA Catalog in European-English.

Paper	Proceeding	Modern Standard Arabic
(White et al., 2008)	Interspeech	✓
(Ali et al., 2021)	Interspeech	✓
(Chowdhury et al., 2021)	Interspeech	✓

Table 22: ISCA Catalog in Middle Eastern-English.

Paper	Proceeding	Hindi	Marathi	Bengali	Telugu	Gujarati	Tamil	Malayalam	Kannada
(Sitaram et al., 2016)	SSW	✓							
(Ramanarayanan and Suendermann-Oeft, 2017)	Interspeech	✓							
(Ganji and Sinha, 2018)	Interspeech	✓							
(Rao et al., 2018)	Interspeech	✓							
(Thomas et al., 2018a)	Interspeech	✓					✓		
(Srivastava and Sitaram, 2018)	Interspeech	✓							
(Rambabu and Gangashetty, 2018)	SLTU	✓							
(Taneja et al., 2019)	Interspeech	✓							
(Rallabandi and Black, 2019)	Interspeech	✓	✓		✓				
(Prakash et al., 2019)	SSW		✓						
(Sharma et al., 2020)	Interspeech	✓							
(Manghat et al., 2020)	Interspeech							✓	
(Bansal et al., 2020b)	Interspeech	✓							
(Chandu and Black, 2020)	Interspeech	✓							
(Kumar et al., 2021)	Interspeech	✓		✓					
(Liu et al., 2021)	Interspeech					✓	✓	✓	
(Diwan et al., 2021)	Interspeech	✓		✓					
(Klejch et al., 2021)	Interspeech	✓		✓					
(Wiesner et al., 2021)	Interspeech	✓		✓					
(Antony et al., 2022)	Interspeech	✓							
(Manghat et al., 2022)	Interspeech							✓	

Table 23: ISCA Catalog in South Asian-English.

Paper	Proceeding	Malay
(Yeong and Tan, 2010)	SLTU	✓
(Yeong and Tan, 2014)	Interspeech	✓
(Singh and Tan, 2018)	Interspeech	✓

Table 24: ISCA Catalog in South East Asian-English.

Paper	Proceeding	Chinese-Taiwanese
(Lyu and Lyu, 2008)	Interspeech	✓

Table 25: ISCA Catalog in Language with Dialects.

Paper	Proceeding	Frisian-Dutch	Russian-Ukrainan
(Lyudovyk and Pylypenko, 2014)	Interspeech		✓
(Yılmaz et al., 2016)	Interspeech	✓	
(Yılmaz et al., 2017b)	Interspeech	✓	
(Yılmaz et al., 2017a)	Interspeech	✓	
(Yılmaz et al., 2018)	Interspeech	✓	
(Yilmaz et al., 2018)	SLTU	✓	
(Wang et al., 2019)	Interspeech		✓
(Yılmaz et al., 2019)	Interspeech		✓

Table 26: ISCA Catalog in Two Languages in the same family.

Paper	Proceeding	Kazakh-Russian	Hindi-Tamil	French-Arabic
		1	1	4
(Amazouz et al., 2017)	Interspeech			✓
(Thomas et al., 2018a)	Interspeech		✓	
(Wottawa et al., 2018)	Interspeech			✓
(Chandu and Black, 2020)	Interspeech			✓
(Chowdhury et al., 2021)	Interspeech			✓
(Mussakhojayeva et al., 2022b)	Interspeech	✓		

Table 27: ISCA Catalog in Two Languages in different families.

Paper	Proceeding	Italian-German-English	Kiswahili-Shen-English
		1	1
(Knill et al., 2020)	Interspeech	✓	
(Otundo and Grice, 2022)	SpeechProsody		✓

Table 28: ISCA Catalog in Trilingual.

Paper	Proceeding	SEA Mandarin-English	African Languages-English	Indian Languages-English	Others
		17	1	1	7
(Badino et al., 2004)	Interspeech				✓
(Oria and Vetek, 2004)	Interspeech				✓
(Marcadet et al., 2005)	Interspeech				✓
(Romsdorfer and Pfister, 2006)	ML				✓
(Lyu et al., 2010b)	Interspeech	✓			
(Imseng et al., 2010)	Interspeech				✓
(Weiner et al., 2012b)	SLTU	✓			
(Adel et al., 2014c)	Interspeech	✓			
(Adel et al., 2014b)	Interspeech	✓			
(Giwa and Davel, 2014)	Interspeech		✓		
(Adel et al., 2014a)	SLTU	✓			
(Rallabandi and Black, 2017)	Interspeech			✓	
(Chandu et al., 2017)	Interspeech				✓
(Garg et al., 2018b)	Interspeech	✓			
(Xu et al., 2018)	Interspeech	✓			
(Guo et al., 2018)	Interspeech	✓			
(Chang et al., 2019)	Interspeech	✓			
(Khassanov et al., 2019)	Interspeech	✓			
(Lee et al., 2019b)	Interspeech	✓			
(Zeng et al., 2019)	Interspeech	✓			
(Hu et al., 2020)	Interspeech	✓			
(Li and Vu, 2020)	Interspeech	✓			
(Zhou et al., 2020)	Interspeech	✓			
(Nekvinda and Dušek, 2020)	Interspeech				✓
(Qiu et al., 2020)	Interspeech	✓			
(Liu et al., 2021)	Interspeech	✓			

Table 29: ISCA Catalog in 4+.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section on page 9

☐ A2. Did you discuss any potential risks of your work?
Not applicable. This is a survey paper. There is no potential negative risk.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1 on page 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.