HEATHER CHRISTINE LENT

NLP ACROSS THE RESOURCE LANDSCAPE: DEVELOPMENT IN CREOLE NLP & EVALUATION IN SEMANTIC PARSING

This thesis has been submitted to the Ph.D. School of The Faculty of Science, University of Copenhagen.

NLP ACROSS THE RESOURCE LANDSCAPE: DEVELOPMENT IN CREOLE NLP & EVALUATION IN SEMANTIC PARSING

HEATHER CHRISTINE LENT

KØBENHAVNS UNIVERSITET

Ph.D. Thesis

July 2022

SUPERVISOR:
Anders Søgaard

ASSESSMENT COMMITTEE:
Desmond Elliott, University of Copenhagen
Katharina Kann, University of Colorado Boulder
Julia Kreutzer, Google Research

AFFILIATION:
Department of Computer Science
Faculty of Science
University of Copenhagen

FUNDING:
This project has received funding from the European Union's Horizon 2020 research and in- novation programme under the Marie Skłodowska-Curie grant agreement No 801199

THESIS SUBMITTED:
July 2022

ABSTRACT

Data availability is the crux on which much research in NLP depends. When a language is highly-resourced, it is possible to train large-scale models that demonstrate impressive performance on a wide array of tasks. Unfortunately, a majority of the world's languages are still lower-resourced, and lack sufficient data (if any data) that such models require for training. As a result, NLP research can look very different in lower- versus higher-resourced settings, and both scenarios come with their own particular challenges. For lower-resourced languages, different methods must be developed for overcoming the constraints of limited data, as well as for leveraging resources from other languages, unless significantly more effort is spent on resource creation. Meanwhile for higher-resourced languages, even when performance metrics are competitive, it can be difficult to evaluate what biases a model has learned from large datasets, and to determine a model's concrete limitations. Thus, in this work, we present a collection of studies from these two different settings of data availability: lower-resourced NLP for Creole languages and higher-resourced semantic parsing evaluation.

The set of studies on NLP for Creoles each expand upon the topic from a different angle. First, we explore a linguistically motivated approach for language modeling of Creoles, utilizing a Distributionally Robust Objective, but ultimately find that this method does not outperform standard Empirical Risk Minimization (Chapter 2). Next, we investigate transfer learning for Creoles, as this is a common approach for leveraging information from high-resourced languages for lower-resourced ones. We find that the typical scenario, whereby cross-lingual learning is achieved by training on a set of languages closely related to the target language, cannot be trivially applied to Creoles (Chapter 3). The final two works within the scope of Creole NLP explore the needs for language technology within Creole-speaking communities, as it is critical that researchers do not assume these needs on behalf of a community (Chapter 4), and discuss approaches, progress, and considerations for creating a multitask benchmark dataset for Creoles, which will allow NLP researchers the opportunity to include these languages within their work, and to further develop Creole NLP (Chapter 5).

For the remainder of this work, we present two studies on approaches for evaluating semantic parsers trained for English. In the first, we introduce a test suite for unit testing of text-to-SQL semantic parsers, in order to identify the true strengths and weaknesses of a model, beyond opaque accuracy metrics; we find that even state-

of-the-art models still struggle with simple SQL operations like selecting columns (Chapter 6). In the second, we test three semantic role labeling parsers for their susceptibility to bias against figurative, non-literal utterances, as such language is common in everyday communication. We find that the parser utilizing a large-scale pre-trained language model was more biased against figurative language, than the models using other word representation approaches (Chapter 7).

ABSTRAKT

Datatilgængelighed er grundstenen som en stor del af forskningen i NLP afhænger af. Når et sprog har adskillige tilgængelige ressourcer, er det muligt at træne modeller der viser en imponerende ydeevne på en bred vifte af opgaver i stor skala. Desværre har et flertal af verdens sprog stadig få ressourcer og mangler tilstrækkelige data (hvis de overhovedet har data), som NLP modeller kræver til træning. Som følge heraf er NLP-forskning markant forskellig i sprog med færre ressourcer, og begge typer sprog kommer med deres egne særlige udfordringer. For sprog med færre ressourcer skal der udvikles anderledes metoder til at overvinde udfordringerne ved begrænset data og der skal udnyttes ressourcer fra andre sprog, medmindre der kommer væsentligt fokus på ressource skabelse i disse sprog. Modsat kan det for sprog med flere ressourcer være vanskelig at evaluere hvilke *biases* en model har lært fra store datasæt, selv når præsentationsparametrene er konkurrencedygtige. Det kan ligeledes være vanskelig at bestemme en models konkrete begrænsninger. I denne afhandling præsenteres en samling undersøgelser fra sprog med to forskellige niveauer af datatilgængelighed: NLP for Kreolsprog med færre ressourcer og evaluering af *semantic parsing* for sprog med flere ressourcer.

Studierne omhandlende NLP for Kreolsprog undersøger emnet fra forskellige vinkler. Først udforskes en sprogligt motiveret tilgang til sprogmodellering af Kreolsprog, ved at benytte et *Distributionally Robust Objective*. Konklusionen er at denne metode ikke er bedre end standard *Empirical Risk Minimization* (kapitel 2). Dernæst undersøges overførselslæring for Kreolsprog, da dette er en almindelig tilgang til at benytte information fra sprog med flere ressourcer til sprog med færre ressourcer. Konklusion er at det typiske scenarie, hvor tværsproglig læring opnås ved at træne modeller på et sæt sprog der er tæt relateret til oprindelige sprog, ikke kan anvendes trivielt på Kreolsprog (kapitel 3). De resterende to studier inden for Kreolsproget NLP undersøger behovene for sprogteknologi i Kreolsprogs-talende samfund, da det er afgørende at forskere ikke påtager sig disse behov

på vegne af et fællesskab (kapitel 4), samt diskuterer tilgange, fremskridt og overvejelser for at skabe et *multitask benchmark-dataset* for Kreolsprog, som giver NLP-forskere mulighed for at inkludere disse sprog i deres arbejde og videreudvikle Kreolsproget NLP (kapitel 5).

I den resterende del af denne afhandling præsenteres to studier om tilgange til evaluering af *semantic parsers* trænet på engelsk. I den første introduceres en *test suite* til modultestning af tekst-til-SQL *semantic parsers*, for at identificere de sande styrker og svagheder ved en model. Ud over uigennemsigtige målinger af nøjagtighed; opdages der at selv avancerede modeller stadig har problemer med simple SQL-kommandoer som at vælge kolonner (kapitel 6). I det andet studie testes tre *semantic role labeling parsers* for deres modtagelighed for *biases* mod figurative, ikke-bogstavelige ytringer, da et sådant sprog er almindeligt i daglig kommunikation. Konklusionen er at *parsers*, der anvender en storstilet præ-trænet sprogmodel, var mere forudindtaget mod figurativt sprog end modellerne, der brugte andre *word representation* tilgange (kapitel 7).

PUBLICATIONS

This is an article-based thesis. The publications below are those included in the thesis. They have been peer-reviewed, and accepted to venues of NLP research.

Lent, Heather, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard (Nov. 2021a). "On Language Models for Creoles." In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 58–71. DOI: 10.18653/v1/2021.conll-1.5. URL: https://aclanthology.org/2021.conll-1.5.

Lent, Heather, Emanuele Bugliarello, and Anders Søgaard (May 2022). "Ancestor-to-Creole Transfer is Not a Walk in the Park." In: *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. Dublin, Ireland: Association for Computational Linguistics, pp. 68–74. DOI: 10.18653/v1/2022.insights-1.9. URL: https://aclanthology.org/2022.insights-1.9.

Lent, Heather, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard (2022). "What a Creole Wants, What a Creole Needs." In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6439–6449. URL: https://aclanthology.org/2022.lrec-1.691.

Lent, Heather and Anders Søgaard (Nov. 2021). "Common Sense Bias in Semantic Role Labeling." In: *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*. Online: Association for Computational Linguistics, pp. 114–119. DOI: 10.18653/v1/2021.wnut-1.14. URL: https://aclanthology.org/2021.wnut-1.14.

Lent, Heather, Semih Yavuz, Tao Yu, Tong Niu, Yingbo Zhou, Dragomir Radev, and Xi Victoria Lin (Nov. 2021b). "Testing Cross-Database Semantic Parsers With Canonical Utterances." In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 73–83. DOI: 10.18653/v1/2021.eval4nlp-1.8. URL: https://aclanthology.org/2021.eval4nlp-1.8.

Additional peer-reviewed publications **not** included in the thesis are listed below. These are publications co-authored by me during the span of the PhD fellowship.

Aralikatte, Rahul, Mostafa Abdou, Heather C. Lent, Daniel Hershcovich, and Anders Søgaard (2020). "Joint Semantic Analysis with Document-Level Cross-Task Coherence Rewards." In: *CoRR* abs/2010.05567. arXiv: 2010.05567. URL: https://arxiv.org/abs/2010.05567.

Aralikatte, Rahul, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard (Nov. 2019). "Rewarding Coreference Resolvers for Being Consistent with World Knowledge." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 1229–1235. DOI: 10.18653/v1/D19-1118. URL: https://aclanthology.org/D19-1118.

Cui, Ruixiang, Rahul Aralikatte, Heather C. Lent, and Daniel Hershcovich (2021). "Multilingual Compositional Wikidata Questions." In: *CoRR* abs/2108.03509. arXiv: 2108.03509. URL: https://arxiv.org/abs/2108.03509.

Hershcovich, Daniel et al. (May 2022). "Challenges and Strategies in Cross-Cultural NLP." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6997–7013. DOI: 10.18653/v1/2022.acl-long.482. URL: https://aclanthology.org/2022.acl-long.482.

Words bounce.
Words, if you let them,
will do what they want to do
and what they have to do.

— Anne Carson

ACKNOWLEDGMENTS

Every PhD is a team effort, and I am extremely grateful to everyone who has been on my team.

Thank you to my wonderful lab mates in CoAStal - the best ever NLP lab! Each of you is kind, cool, fun, and brilliant.

Thank you to Emanuele, my CoAStal parter in crime. I would have gone crazy without our afternoon tea breaks, as well as all of your encouragement, and help with latex.

Thank you to my supervisor, Anders, for being a great supervisor, encouraging me through a PhD's ups and downs, and for always knowing how to put a positive spin on something.

Thank you to my Study Buddies around the world, who kept me focused and made working during the pandemic a lot more fun.

Thank you to all my friends and family, who celebrated my successes with me, supported me through the hard times, and listened to me complain about my various projects for three years.

Thank you to Christoffer, who brought me countless snacks and cooked me many meals during big deadlines, and always cheered me on.

Mèsi anpil to Michel DeGraff, for patiently answering all of my questions, and for helping me to both learn, and unlearn, a lot about Creoles.

Thank you to the people who invested in me, early in my academic career, as I wouldn't have had the chance to do a PhD without you: 감사합니다 to Hyun Seok Park, and thank you to Eric Lyons and Mihai Surdeanu.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Part I

BACKGROUND

1 INTRODUCTION

Natural language processing (NLP) is largely contingent upon data availability, and as a result, research often looks very different for higher-resourced languages than it does for lower-resourced ones. For instance, while it is possible to train large-scale models composed of millions of parameters for highly-resourced languages, NLP for lower-resourced languages often hinges on methods for linguistic transfer, as little or no data are available. And while researchers drive progress on complex, high-level tasks like multi-hop question answering or text-to-SQL semantic parsing for higher-resourced languages, their counterparts working in a lower-resourced setting are typically preoccupied with more fundamental, low-level tasks like part-of-speech tagging and dependency parsing – technologies that are largely taken for granted as being reliably available and generally accurate, for higher-resourced languages, like English. And finally, while researchers develop better methods for evaluating the capabilities, deficiencies, and biases that models learn from massive, enigmatic corpora like Common Crawl[1], evaluation within lower-resourced NLP may often come with fewer surprises, as smaller datasets can often allow researchers the chance to become closely acquainted with their training data, or even to perform a thorough qualitative error analysis by hand.

Indeed, the divergence between research in higher- versus lower-resourced settings demonstrates how NLP as a field can at times be very disjointed – and this is something to be wary of. After all, estrangement of the NLP community into different research silos can greatly hamper progress, as important lessons learned in one setting risk not being communicated to the other side, and thus effectively being "lost". In accordance, the most efficient path towards the end goal of NLP (i.e. high-performing language technologies available for all languages that want them) will require exchange of knowledge from research focused on both higher- and lower-resourced settings. Thus, in this thesis, we present a unified exploration of NLP, across the data availability landscape, including studies on lower-resourced NLP for Creole languages (Part II) and evaluation of higher-resourced semantic parsers (Part III). In presenting these two ostensibly disconnected topics, from opposite ends of the data availability spectrum, we hope to encourage more conversation and cohesion within the field of NLP.

1 https://commoncrawl.org/

1.1 LOW-RESOURCE NLP FOR CREOLES

Creole languages (or, simply, *Creoles*) are a diverse set of languages, that can be found all over the world. Some examples of Creoles include Haitian Creole (Caribbean), Sranan Tongo (South America), Louisiana Creole (North America), Nigerian Pidgin English (Africa), Singaporean Colloquial English (Asia), Australian Krio (Australia), and Tok Pisin (Pacific). Despite being scattered globally, what these languages share in common is having a history of language evolution, which involved linguistic contact between multiple, *unrelated* languages. Haitian Creole, for instance, has notable influences from French, Fongbe, and Igbo (DeGraff, 2007). And although some Creole languages can boast numbers of speakers in the millions, they remain lower-resourced languages in NLP, with some Creoles having no resources at all.

LINGUISTIC BACKGROUND Creoles are a particularly contentious set of languages within the field of linguistics, and have been so for at least half a century (Alleyne, 1971; Bickerton, 1984; DeGraff, 2001, 2003, 2005b; McWhorter, 1998; Muysken and Smith, 1986; Parkvall et al., 2008; Sessarego, 2020). Debates on Creoles are typically driven by two questions: (1) "How did Creoles originate?" (i.e., Creole genesis) and, (2) "Are Creoles different from other languages?". The first is a question of language evolution, within the context of specific historical events: for many (but not all) Creoles, and particularly Caribbean Creoles, the origins of the language are inseparably tied to European colonization and the Atlantic slave trade [2]. That is to say, the reason why European languages (i.e., Portuguese, Spanish, Dutch, French, and English) came into contact with unrelated African languages (e.g. Akan, Fon, Igbo, Hausa, Yoruba, etc.), was because European colonizers forcibly enslaved and displaced Africans. Linguists do not debate that Creoles are the result of linguistic contact between these languages, but rather debate the process by which a Creole *became a brand new language*. The most commonly purported theory for Creole genesis is the pidgin-to-Creole hypothesis, which suggests that when multiple, unintelligible languages come into contact with one another, a simplified *pidgin* language is developed by the community to facilitate communication; then as children in the community learn the pidgin as a native language, the pidgin undergoes creolization, and thereby becomes a full-fledged Creole language (Kouwenberg and Singler, 2009). Although this theory is certainly prevalent across linguistic literature, it is not without its opponents. For example, the pidgin-to-Creole hypothesis has been harshly criticized by

2 There are also Creoles with little to no influence from European languages. Some examples include Lingala and Kikongo-Kituba, heavily influenced by Bantu; Juba Arabic and Kinubi influenced by Arabic; and Sri Lankan Malay influenced by Malay. See (Michaelis et al., 2013) for more information.

Aboh (2015), because it can be viewed as contradicting aspects core to the notion of Universal Grammar (Chomsky and Lasnik, 2008): if all humans should share *equally* an innate capacity for language, why should Creole evolution be any different from language evolution across other languages? The claim that Creoles formed differently from other languages, implies that either the original speakers of Creole languages (who were often enslaved Africans) had less capacity for language – an obviously incorrect and racist conclusion – or that some linguistic universals do not, in fact, apply to all languages (i.e. a contradiction of linguistic universals) (Aboh, 2015).

Thus, we can see that this first question on Creole genesis naturally begets the second question on Creole uniqueness: if the process by which Creoles originated was indeed somehow distinct from the typical language evolution of other languages, could this somehow make Creoles themselves linguistically different from all other languages? Proponents of the idea that Creoles are indeed *exceptional* from other languages typically point to examples within Creole grammars that seem simplified when compared to the original languages that influenced it, as well as that Creoles seem to exhibit limited morphological complexity (Bickerton, 1984; McWhorter, 1998). Meanwhile opponents to the idea of Creole Exceptionalism point to instances where Creole grammar appears to be more complex than the other relevant languages, as well as examples of Creole utterances displaying complex morphology (DeGraff, 2003; Henri, Stump, and Tribout, 2020).

While these linguistic discussions are interesting and relevant, (as we will see in Chapter 4), these linguistic debates should not impede on efforts to develop language technology for those Creole-speaking communities that desire such technologies. At the end of the day, there are real communities speaking these languages, as well as people who can call a Creole their mother tongue.

RELATED WORK Presently, there are a very limited number of works exploring Creoles directly, within the scope of computational linguistics and NLP. Published studies in computational linguistics on Creoles utilize common methods from population genetics to further expand on the linguistic debates of Creole genesis and Creole Exceptionalism, but perhaps unsurprisingly, these studies do not yield a consensus regarding these debates (Daval-Markussen and Bakker, 2012; Murawaki, 2016). Meanwhile, the remaining bulk of published work exploring Creole NLP directly are included in this thesis (Chapters 2, 3, and 4). Readers can also refer to our survey of related works within computational linguistics and NLP for Creoles in Chapter 4, which includes an in-depth discussion of Daval-Markussen and Bakker (2012) and Murawaki (2016), as well as an audit of published resources for Creole NLP.

CONTRIBUTIONS The works presented in this thesis broaden existing knowledge on Creole NLP in a number of ways. Chapter 2 provides the first ever investigation of a tailored approach to language modeling for Creoles, inspired by the multilingual nature of Creole lexicons, as well as the social dynamics common amongst speakers of some Creole languages. We employ a Distributionally Robust Objective, to encourage the model to be robust to multilingual vocabulary, however, this approach does not perform better than typical Empirical Risk Minimization. We found that this may be due to relative stability within the Creoles, or at least stability within our Creole data. Next, Chapter 3 explores the limitations of transfer learning for Creoles, demonstrating that cross-lingual learning for Creoles is not immediately as trivial as for other, non-Creole languages. Furthermore, based on some unique, but consistent, behaviors we observe during multilingual training for Creole languages, we test whether training for long (i.e., overfitting on the set of languages related to the Creole), can lead to better language models for Creoles, but ultimately discredit this hypothesis. In Chapter 4, we take a step back from developing NLP technologies for Creoles, to engage with Creole-speaking communities, and underscore the importance of community involvement, when developing NLP for lower-resourced languages. This work also contributes a survey of existing works in computational linguistics and NLP for Creoles, including documentation of available and unavailable datasets for Creoles. Finally, in Chapter 5, we provide a detailed discussion of our ongoing work to create a benchmark dataset for Creoles. We note that Chapter 5 is the only work *not* peer reviewed in this thesis, and it has not been submitted to a conference or journal, as development of the dataset is still underway. That said, this chapter still provides marked contributions to the thesis, such as detailed discussion about data creation, an investigation into common issues with Wikipedia, the presentation of a brand new cross-cultural NLP task, and a precise plan for finalizing the dataset.

CORRECTIONS & CLARIFICATIONS In the published manuscript for Chapter 2, the "C" in "Creole" was originally lower-cased, and this has been corrected for this thesis. Although the word "Creole" is often written as lower-cased "creole" in academic literature, and even by creolists themselves, DeGraff (2020) has pointed how the discrepancy in capitalizing other respected language groups, (e.g., "Germanic", "Romance", etc.), while maintaining a lower-cased "creole", reinforces existing inequalities between Creoles and other groups of languages. Moreover, in Chapter 2, we present the commonplace pidgin-to-Creole hypothesis, without context that this is just *one* theory of Creole genesis. Chapter 2 has *not* been modified to remove this theory, as subsequent Chapters 3 and 4 correct this mistake, and present

a more neutral description of Creoles, with more careful attention to different theories about their origins.

It is also worth clarifying that, while the study presented in Chapter 3 observes unique pattern of behavior for Creoles in contrast with the non-Creoles, this thesis does not take a side in the linguistic debate regarding Creole Exceptionalism (i.e., the notion that Creoles form a unique language class). Readers interested in exploring this topic further can find many linguistic works, discussing this debate directly (DeGraff, 2005b; Migge, 2020; Mufwene, 2014).

1.2 HIGH-RESOURCE SEMANTIC PARSING EVALUATION

Semantic parsing is a popular task within NLP, with the goal of automatically mapping natural language utterances to logical forms that capture the utterance's semantic meaning. The type of logical form can vary in practice, for example, some are *graphs* motivated by linguistic frameworks, such as frame semantics (Ringgaard, Gupta, and Pereira, 2017), semantic role labeling (Marcheggiani and Titov, 2017), AMR (Banarescu et al., 2013), or UCCA (Hershcovich et al., 2019); others are *programming languages* like SQL (Dahl et al., 1994) or Python (Yin et al., 2018), which encode the meaning into executable commands. Across all of these flavors of semantic parsing, however, most datasets still exist only in English, arguably the highest-resourced language in all of NLP (Joshi et al., 2020b). And although some semantic parsing datasets may be small (e.g. Geoquery with 880 examples (Iyer et al., 2017)), many datasets within semantic parsing still contain a few ten-thousand examples (Damonte and Monti, 2021). While more data would undoubtedly be beneficial, we must remember that these English language semantic parsing datasets are still much larger, and more diverse, than even the highest quality datasets for low-resource languages. For example, the Spider dataset (Yu et al., 2018) for text-to-SQL semantic parsing includes over 10,181 natural language questions, spanning 200 databases covering 128 domains. Compare this against the MasakaNER dataset for Nigerian Pidgin (Adelani et al., 2021), with 3,000 examples in total, covering just 1 domain. Moreover, even for smaller semantic parsing datasets, it is still possible to leverage available tools, previously trained on massive amounts of English text, such as language models (Devlin et al., 2019; Liu et al., 2019) and part-of-speech taggers (Bohnet et al., 2018). Therefore we argue that English semantic parsing is a relatively high-resourced task within NLP as a field.

EVALUATION NLP models need evaluation to judge their efficacy, whether they were trained on a little data, or a lot. And yet, it has been well established that evaluating models, especially within the scope of high-resourced NLP, can be very difficult. To start, held-out

test sets only provide a limited view into model performance, but still cannot indicate whether a model is truly ready for deployment to end-users (Ribeiro et al., 2020). To overcome this limitation, several works have contributed approaches for more comprehensive evaluation of NLP models, such as looking at logical consistency (Elazar et al., 2021) or robustness to adversarial examples (Iyyer et al., 2018). These evaluation frameworks are useful, especially in contexts when we suspect a model may be brittle along those dimensions, due to some shortcomings in the training data. Though, as Bender et al. (2021) point out, anticipating model weaknesses stemming from shortcomings within a dataset, requires that the dataset, and its flaws, must be *knowable*, which is certainly not possible for extremely large-scale datasets. As a consequence, models trained on such datasets (or utilizing other NLP components trained on such datasets, like pre-trained language models) introduce a further challenge to evaluation: there may be model weaknesses or biases that need evaluation, but we don't know about them (i.e., we don't know what we don't know, and cannot evaluate what we don't know). Thus, evaluation of high-resource NLP systems remains an open problem. Moreover, it should be noted that evaluation of models is also not equally straight forward across different tasks within NLP – some tasks are naturally easier for humans to make quick judgements about. For example, it is much easier for a human to evaluate the predictions of a sentiment analysis model (i.e., determine if some text has a positive or negative sentiment, and compare this to the model's classification), than the predictions of a semantic parser (i.e., determine whether a model's predicted parse matches the meaning of the sentence). The latter also requires a trained expert, knowledgeable about a parser's output formalism, whether that be UCCA graphs or SQL queries, which are much more tedious and complicated to evaluate. Consequently, much work on model evaluation shies away from more difficult tasks like semantic parsing (Ribeiro et al., 2020), and more works aiming to improve evaluation methods for semantic parsing are greatly needed.

TEXT-TO-SQL SQL is a popular programming language used to query relational databases. Intuitively, text-to-SQL semantic parsing can be thought of as translation of natural language utterances (e.g. questions or commands) into executable SQL queries, which will be run against a database. If the predicted SQL query is correct, then the question will be answered, with knowledge from the database. Therefore, a high quality text-to-SQL system, should allow anybody to access information within a database, without requiring them to be proficient in SQL.

SEMANTIC ROLE LABELING The notion of semantic roles was first introduced by Gruber (1965), and has become an important frame-

work within modern linguistics for exploring the relationship between syntax and semantics. Within any sentence, different words and phrases will play a different role. For example, in the sentence: *"Lindsay gave George Michael the ring"*, the entities "Lindsay", "George Michael", and "ring", each entail a different semantic meaning within the sentence. Lindsay can be said to be the *agent* of the sentence, as she is the one taking action, by giving a ring; Lindsay is also the subject of the sentence, as *agents* tend to be. Meanwhile, George Michael can be understood as the *recipient* within the sentence, as he receives the ring. As the action of giving typically requires specification of *to whom* the giving has occurred, it is natural that the *recipient* semantic role is often also the sentence's indirect object. Finally, the ring can be understood as the *theme* of the utterance, as it is the thing transferred from Lindsay to George Michael. The semantic role of *theme* can also often occupy the syntactic role of direct object. Aside from this simple example, though, many more semantic roles exist within the traditional linguistic framework (e.g. *experiencer*, *patient*, and *location*) (Bornkessel et al., 2009).

While this explanation demonstrates the origins and intuition behind semantic role labeling, it must be noted that in practice, existing semantic role labeling datasets do not do classification of *agents* and *themes*, as such. Rather, prominent semantic role labeling datasets, such as OntoNotes (Hovy et al., 2006a) set up the task as identification of *verbs* (predicates) and their associated *arguments* and potential *modifiers*. For instance, if we know the verb "give" requires three arguments (i.e. *who-0* gave *what-1* to *whom-2*), we can thus re-imagine our example sentence in the context of OntoNotes, where a correct semantic parse would identify the verb, *gave*-V, and its arguments *Lindsay*-ARG0, *book*-ARG1, and *Geroge Michael*-ARG2. If we changed our sentence slightly to have, *"Lindsay will soon give the George Michael the ring"*, then we would additionally have labels for *will*-ARGM-MOD and *soon*-ARGM-TMP, as these words modify the anchor verb *give*-V, with additional information about tense and time. Thus, the goal of semantic role labeling in NLP is to train a parser, capable of automatically identifying these verbs (predicates), arguments, and modifiers; for a parser to successfully do this, it must understand the meaning of a given verb, to determine how many arguments it can have and which words can possibly modify it, as well as how these parts all relate to each other syntactically, so as not to mislabel an irrelevant non-argument.

CONTRIBUTIONS The studies presented in Chapters 6 and 7 both demonstrate different methods for fine-grained evaluation of different semantic parsers. In Chapter 6 we introduce an approach for generating high quality data, which is used to perform unit testing on state-of-the-art text-to-SQL models. The unit tests are designed such

that only one specific SQL element is being tested within a single test example. Thus, we are able to perform a low-level evaluation of the parsers, evaluating them on fundamental skills like selecting columns, understanding logical operators, and so on. In the end, we find that even when a text-to-SQL parser has a high accuracy on the held-out test set, it may still struggle with very basic SQL operations, like joining two columns. Thus, our evaluation framework provides a much more in-depth look into the strengths and weaknesses of a parser, than simple accuracy over a test set. Next, in Chapter 7, we evaluate three semantic role labeling parsers on their susceptibility to *common sense bias*. A parser can be understood to be biased towards common sense utterances, if it performs worse on figurative language, than literal language. Consider the following two examples: (1) "John body-slammed Tom", and (2) "Love body-slammed Tom". Although these sentences share the exact same syntactic structure, the semantic roles of "John", "Love" are very different in the first and second examples, as the first can be interpreted as a literal utterance, but the second would typically be understood as a figurative utterance, unless there was a person named "Love". Ideally, a parser will be able to correctly identify the differing semantic roles within both sentences, despite the syntax being the same. To test this, we generate similar examples of simple transitive sentences, and test the parsers' performance over them. In the end, we find that the semantic role labelers using large, pre-trained language models, are more biased against figurative language, and therefore exhibit the common sense bias. Meanwhile, the parsers utilising other word representation methods were more robust to figurative language. We also contribute a dataset for testing SRL systems for common sense bias.

Part II

CREOLE NLP

2

ON LANGUAGE MODELS FOR CREOLES

2.1 ABSTRACT

Creole languages such as Nigerian Pidgin English and Haitian Creole are under-resourced and largely ignored in the NLP literature. Creoles typically result from the fusion of a foreign language with multiple local languages, and what grammatical and lexical features are transferred to the Creole is a complex process (Sessarego, 2020). While Creoles are generally stable, the prominence of some features may be much stronger with certain demographics or in some linguistic situations (Patrick, 1999; Winford, 1999). This paper makes several contributions: We collect existing corpora and release models for Haitian Creole, Nigerian Pidgin English, and Singaporean Colloquial English. We evaluate these models on intrinsic and extrinsic tasks. Motivated by the above literature, we compare standard language models with distributionally robust ones and find that, somewhat surprisingly, the standard language models are superior to the distributionally robust ones. We investigate whether this is an effect of over-parameterization or relative distributional stability, and find that the difference persists in the absence of over-parameterization, and that drift is limited, confirming the relative stability of Creole languages.

2.2 INTRODUCTION

A Creole language arises if a *pidgin*,[1] developed by adults for use as a second language, becomes the native and primary language of their children. Although a large portion of Creole languages have their roots in Western European colonialism and slavery, Creole languages still serve as important *lingua franca* in multi-ethnic and multilingual communities, and Creoles are often an important part of the local identity. Moreover, there are more than a hundred million speakers of Creole languages world wide (Figure 1), with similar needs for technological assistance, and yet Creoles are still largely absent from NLP research (Joshi et al., 2020b). Haitian Creole, for example, has 9.6 million speakers as of today; Nigerian Pidgin English has 100 million speakers, and Singaporean Colloquial English (Singlish) has 3.5 million speakers. This paper sets out to collect existing resources

1 A pidgin is a grammatically simplified language that develops between two or more groups that do not have a language in common. Both pidgins and Creoles are sometimes referred to as *contact* languages.

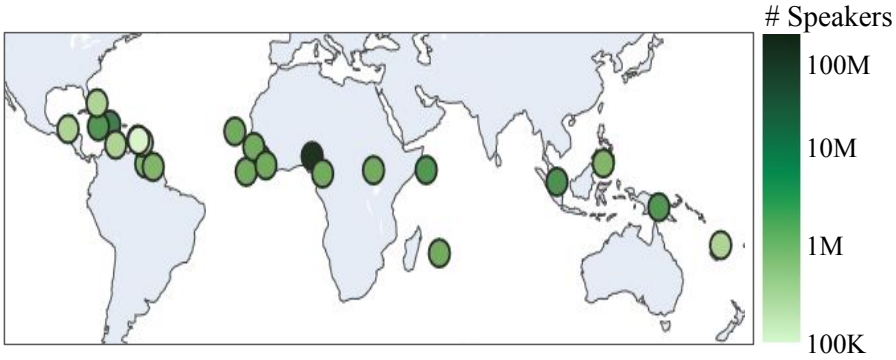


Figure 1: Creoles with a minimum of a hundred thousand speakers are shown here (Hawaiian Pidgin not pictured). Approximately 180 million Creole speakers are represented in this map. Data extracted from https://en.wikipedia.org/wiki/List_of_creole_languages.

for these three languages and provides language models for them. In doing so, we wish to take the nature of Creole languages into account, not necessarily assuming that our best approaches to modeling non-Creole language are also best for the Creole languages.

The nature of Creole languages has been a matter of much debate in linguistics during the last decade (Sessarego, 2020): Some see Creole languages as natural stages in language change cycles (Aboh, 2015), while others see them as a distinct typological class with unique characteristics, including, for example, a very simple morphology (McWhorter, 1998). Another feature of Creoles is that they exhibit significant variation across groups of speakers (Patrick, 1999). Winford (1999) goes as far as to call Creoles a *continua that cannot be captured under a single grammar*.

Consider the following pair of sentences from Bajpai et al. (2017):

(1) John sibei hum sup one.

(2) John very buaya sia.

Here, according to the authors, both sentences are valid utterances in Singlish, and they both mean *John is so lecherous*, but the first would more likely come from a speaker of Chinese, and the second from a Malay speaker. From this,[2] we derive the conjecture that Creole language models can benefit from learned mixtures of source languages. Training on mixtures of source languages has been applied to language modeling of code-switched language (Pratapa et al., 2018),

2 Creole languages clearly differ though in the dynamics that affect their drift. For example, Yakpo (2021) discuss two seemingly similar Creole languages, Krio (Sierra Leone) and Pichi (Equatorial Guinea). Both Creoles have English as their lexifier, but while Krio is spoken alongside English, Pichi is spoken alongside Spanish. The two Creoles, as a consequence, exhibit a clear difference. Krio has converged increasingly toward English, while Pichi has neither converged toward English nor Spanish.

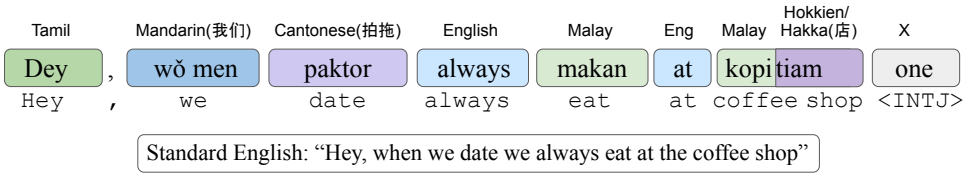


Figure 2: Example sentence in Singlish featuring multilingual vocabulary, Chinese-style topic prominence combined with a subordinate clause with English word order, and a final interjection representing a discourse particle; a common feature of Singlish. Example from https://languagelog.ldc.upenn.edu/nll/?p=25758.

and it is clear from examples such as the one in Figure 2 that Creole languages, at the sentence level, share commonalities with code-switched language, with vocabularies drawn from multiple source languages. To exploit synergies with learned mixtures of source languages, and to obtain robust performance across related, but unseen distributions, we explore ways of training Creole language models with distributionally robust objectives (Oren et al., 2019a). Our results below, however, show that, somewhat surprisingly, this conjecture is probably not true, at least not in a straight-forward way.

CONTRIBUTIONS We combine existing datasets and present pre-trained language models for the following Creole languages: Nigerian Pidgin English, Singaporean Colloquial English (Singlish), and Haitian Creole. We perform intrinsic evaluation (word prediction), as well as extrinsic evaluation (part-of-speech tagging and named entity recognition). Comparing language models trained with empirical risk minimization to languages models trained with robust objectives, we observe that training with multiple related languages does not improve Creole modeling; and also, somewhat surprisingly, that models with empirical risk minimization are superior to models robust across domains. We hence investigate *why* this is: in particular, whether it is due to over-parameterization, insufficient regularization (Sagawa et al., 2019), or relative distributional stability (Ben-David et al., 2007). We observe no significant difference for language models with fewer parameters or higher degree of regularization. On the other hand, we find that the underlying reason might be the relative stability of the Creoles, which show no significant drift.

2.3 RELATED WORK

NLP RESEARCH ON CREOLES Despite the unique features of Creoles that make them an interesting application for multilingual and cross-lingual NLP, as well as the open-ended debate about the linguistic nature of Creoles (Sessarego, 2020), little attention has been

devoted to Creoles in NLP. (We present the works related to the specific Creoles of focus in this paper in §2.4.)

One relevant work by Murawaki (2016) explored the typological status of Creoles and also introduced a method for statistical modeling of Creole genesis. To start, the authors reported that binary SVM classification of Creole and non-Creole languages failed to distinguish the two classes, even though their underlying distributions are quite different. After this, they introduce a statistical model of Creoles, formulated as a mixture of its influential languages and an inferred "restructurer", which is set of possible linguistic feature distributions that are observed across languages included in their experiments. Overall, this work showcases how statistical modeling methods can be useful for investigating the language evolution of Creoles, however there is also no discussion of how their findings could help others extend current NLP methods for Creoles.

NLP RESEARCH ON PIDGINS AND CODE-SWITCHING Creoles are pidgins that have consolidated over time to become a first language for new generations of speakers. The NLP literature on pidgins is even more sparse than the literature on Creoles, because many pidgins that did not undergo creolization have gone extinct, such as Maritime Polynesian Pidgin (Kriegel, 2016). Code-switching literature, however, is also relevant, as both pidgins and Creoles also draw from other languages. Importantly, pidgins differ from code-switching or mixed language in that code-switching typically only occurs between two bilingual or highly proficient speakers of two languages. Pidgins, on the other hand, are derived from multiple languages, and spoken by those who do not fluently speak every language involved. The NLP literature on code-switching is surprisingly rich, however. We refer readers to Çetinoğlu, Schulz, and Vu (2016) and Doğruöz et al. (2021) for an overview.

COMPUTATIONAL RESEARCH ON LANGUAGE EVOLUTION Research on Creoles is more common in the field of language evolution than in NLP. In particular, work on Creoles in this field typically focuses on their computational modeling, their emergence (Nakamura, Hashimoto, and Tojo, 2009), and their evolution (Furman and Nitschke, 2020; Jansson, Parkvall, and Strimling, 2015). Other Creole modeling efforts in this space may be more tailored towards specific linguistic insights (Parkvall, 2008). While these studies demonstrate that work on Creoles is being done in a computational space, it is difficult to apply conclusions from them to NLP, because distinct empirical assumptions are made in these two research areas.

DISTRIBUTIONALLY ROBUST OPTIMIZATION Effectively learning to model and predict underrepresented subdistributions has always

been a challenge in machine learning, e.g., when predicting rare classes, (Fei and Liu, 2016; Scheirer et al., 2013) or classes of examples from rare domains (Zheng, Chen, and Huang, 2020) or minority groups (Hashimoto et al., 2018). Often, underrepresented data is ignored or learned poorly by the models (Feldman and Zhang, 2020), compared to their over-represented counterparts. Distributionally Robust Optimization (DRO) (Hashimoto et al., 2018; Sagawa et al., 2019) aims to minimize the loss on *all* sub-populations, rather than minimizing their average (Ben-Tal et al., 2013). DRO has been particularly useful in the domain of algorithmic fairness (Hashimoto et al., 2018), but has also been found to boost performance on underrepresented domains in language modeling (Oren et al., 2019a) and is generally applicable in situations with drift (Koh et al., 2021).

2.4 CREOLES AND CORPORA

While Creole languages are spoken by hundreds of millions, and are often a *lingua franca* within a larger community, only a handful of resources exist for Creoles presently. Some challenges to collecting data resources for Creole languages can be a Creole's non-standardized orthography, e.g. Haitian Creole (Hewavitharana et al., 2011), or the specific contexts in which Creoles are used – it may not always be used in official capacities for news, education, and official documents, even if the Creoles are widely used in most other aspects of life (Shah-Sanghavi, 2017). This of course complicates data collection. In this work, we focus on the following Creoles, as they each have diverse linguistic makeup and have *some* existing datasets:

NIGERIAN PIDGIN ENGLISH West Africa is one of the world's most linguistically diverse places, with Nigeria alone having over 400 languages (Ufomata, 1999). Recent work to advance African NLP has led to the creation of several datasets in Nigerian Pidgin English (Adelani et al., 2021; Agić and Vulić, 2019; Caron et al., 2019; Ndubuisi-Obi, Ghosh, and Jurgens, 2019a; Ogueji and Ahia, 2019; Oyewusi, Adekanmbi, and Akinsande, 2020; Oyewusi et al., 2021b), which makes it particularly well-resourced in comparison to other Creole languages. Nigerian Pidgin English, also referred to as simply Nigerian Pidgin, can further be understood as a member in the larger family of West African Pidgins, as many West African countries have their own unique variation of this Creole, but all share influences from many of the same languages, such as Igbo, Hausa, and Yoruba.

The first sizeable Nigerian Pidgin dataset comes from Agić and Vulić (2019), who collected parallel text from several magazines written by a religious society, which have parallel translations in many languages. This dataset has been utilized in the first attempts to develop baselines for machine translation of Nigerian Pidgin English

(Ahia and Ogueji, 2020; Ogueji and Ahia, 2019). Furthermore, Ogueji and Ahia (2019) also introduced the first corpus of Nigerian Pidgin English to further facilitate machine translation from Nigerian Pidgin into English. Ndubuisi-Obi, Ghosh, and Jurgens (2019a) also introduced a code-switching corpus of news articles and online comments in both Nigerian Standard English and Nigerian Pidgin. In this work, they discuss some challenges of working with Nigerian Pidgin, such as non-standardized spelling. They also find that different topics prompt code-switching to Nigerian Pidgin over Nigerian Standard English. More task-specific Nigerian Pidgin datasets have been introduced for Universal Dependency Parsing (Caron et al., 2019), named entity recognition (Adelani et al., 2021; Oyewusi et al., 2021b), sentiment analysis (Oyewusi, Adekanmbi, and Akinsande, 2020), and speech recognition (Ajisafe et al., 2020; Bigi, Caron, and Abiola, 2017).

SINGLISH Singaporean Colloquial English, also known as *Singlish*, has English as a source language, but also draws parts of its grammar and vocabulary from languages such as Mandarin, Cantonese, Hakka, Hokkien, Malay, and Tamil. Presently, few publicly available datasets exist in Singlish, as this Creole is primarily utilized for informal conversation between people and not for official purposes. The largest relevant corpus is The National University of Singapore SMS Corpus from Chen and Min-Yen (2015a), which consists of over 67,000 text messages written by Singaporeans. Qualitatively, we observed that this dataset is much closer to Standard English, albeit with noise from outdated SMS language, than the example provided in Figure 2, but, within this data, we still observe many hallmark features of Singlish such as discourse markers and vocabulary from relevant languages. Tan et al. (2020) have also released a webcrawler that collects posts from an popular Singaporean forum about hardware, where discussion is often in Singlish. They use the resulting Singlish corpus as part of their work to investigate the role of inflection for NLP with non-standard forms of English. Beyond plain text corpora, Wang et al. (2017) introduced the first Singlish Universal Dependency dataset, which was further expanded upon in Wang, Yang, and Zhang (2019). Chau, Lin, and Smith (2020) used this dataset as a low-resource language test case for their method of pretraining mBERT (Devlin et al., 2019). Finally, a few studies have been done on private datasets for sentiment analysis (Bajpai et al., 2017; Ho et al., 2018a), and polarity detection (Lo et al., 2016).

HAITIAN CREOLE Haitian Creole exhibits a combination of French with many West African languages (e.g. Igbo, Yoruba, Fon, etc.). Haitian Creole seized the attention of the machine translation community in the aftermath of the 2010 earthquake crisis in Haiti, during which Munro (2010, 2013) developed the Haitian Disaster Response Corpus.

Language	Source	Domain
en, fr, es, pt, yo, zh, ta	WMT-News 2020	news
ms	Malay 30k News	news
Nigerian Pidgin	PidginUNMT Corpus	news
Singlish	Singapore SMS Corpus	sms
Haitian Creole	Disaster Response Corpus	sms

Table 1: Data resources utilized in our experiments.

This is a parallel Haitian–English dataset of SMS messages related to the crisis, to enable rapid development of machine translation systems to assist the crisis response. This dataset was included in the 2011 Workshop for Machine Translation (Callison-Burch et al., 2011), in conjunction with data from the medical domain, newswire, and a Haitian glossary.[3] Several studies used this data to extend methods in statistical machine translation (Hu et al., 2011a,b; R. Costa-jussà and Banchs, 2011) as well as spell checking and data cleaning (Stymne, 2011).

2.5 DATASETS FOR CREOLE LANGUAGE MODELS

We experiment with training language models for Creoles with a mixture of Creole data, and additional data from languages influential to each Creole.

DATA SPLITS We begin with the Creole datasets noted in Table 1, and combine them with data of other higher-resource languages that have been influential to the Creole. We combine a fixed number of these examples into a MIXED-LANGUAGE dataset, as described in Table 2. The MIXED-LANGUAGE dataset for each Creole includes information about the original language of each sentence, so that we can form language-specific groups for DRO (see subsec:dro/training for more details on DRO grouping). The total number of train and development examples were determined by the number of sentences in the base (Creole) dataset for a 95-5 train-development split. Singlish had equal representation of each language, with 53,006 examples per language, including Singlish. Haitian Creole also had equally represented languages, with 8,192 examples for Haitian and each additional language. For the Nigerian Pidgin MIXED-LANGUAGE dataset, English, Portuguese, and Nigerian Pidgin were composed equally with 67,615 examples each, and Yoruba with only 27,260 examples

3 http://www.speech.cs.cmu.edu/haitian/text/.

Creole	Langs	# Train Mixed-Lang	# Train Creole-Only	# Dev Creole-Only
Nigerian Pidgin	en, pt, yo	230,105	53,006	3,359
Singlish	en, zh, ms, ta	265,030	67,615	2,790
Haitian Creole	fr, yo, es	32,768	8,192	988

Table 2: Creoles, their influential languages (Langs), and the number of examples in the Train-Dev split for our MIXED-LANGUAGE and CREOLE-ONLY experiments. Both use the same Creole-only dev dataset.

due to the small size of the original data. Thus, we included 95% of the Yoruba WMT-News 2020 dataset.

LANGUAGE IDENTIFICATION WITHIN CREOLES As we will see in §2.7, training the language models on the MIXED-LANGUAGE dataset with DRO fails to produce positive results. Following from this, we also create a CREOLE-ONLY dataset, composed of only the Creole examples. In order to sort the Creole examples into distinct groups for DRO, we label each Creole example by the *collection* of the selected languages present in the sentences, as determined by a language identification algorithm.[4] Consider the following examples from their respective CREOLE-ONLY datasets:

Singlish: *"treat him makah lah"*
en: 88.19%, ms: 4.34%, ta: 0.04%, and zh: 0.01%

Nigerian Pidgin: *"Pikin wey like to play wit wetin no dey common and sabi one particular subject reach ground"*
en: 87.46%, pt: 0.23%, and yo: 0.03%

Haitian Creole: *"Infomation sou kestion te tranble a ak lekol"*
fr: 3.50%, es: 0.08%, and yo: 0.01%

While the language identification algorithm is not perfect, the confidence scores for the languages still reflect the high-level trends for the Creole examples, namely, that English and Malay (*"makan"*) are indeed present in the Singlish sample, and also that English and Portuguese (*"pikin"*, *"sabi"*) are present in the Nigerian Pidgin example. However, for the Haitian Creole example, we see that none of our chosen languages have very high scores from the language identification algorithm, which begs the question: were there other languages with higher confidence from the language identification algorithm?

4 https://fasttext.cc/blog/2017/10/02/blog-post.html.

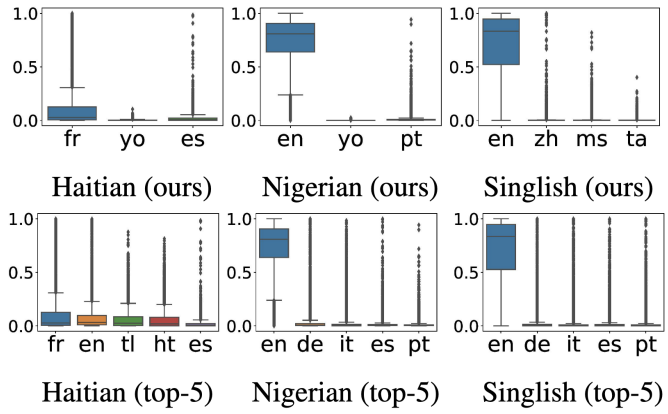


Figure 3: Distributions of identified languages across the CREOLE-ONLY test set. **Top:** distributions for the influential languages included in MIXED-LANGUAGE. **Bottom:** distributions of the five languages that had the highest prediction scores for each Creole, where we see a bias towards European languages.

To ensure that the languages we chose are well-represented in the Creole examples, we looked at the distribution of the identified languages across examples in our CREOLE-ONLY datasets in Figure 3. From this, we observe that choosing to identify languages specifically related to the Creole (i.e. the same languages we included in the MIXED-LANGUAGE datasets) is more reliable than trusting the language identifier pick the top five languages with the highest confidence – there appears to be a bias for falsely predicting European languages, even on Creole data unrelated to these languages, as well as some strange outliers, such as Tagalog being the third most commonly predicted language for Haitian Creole sentences. Also, we see that Haitian Creole itself was a commonly identified language, which could explain the low confidence scores for French and Spanish in the example above. Finally, among our specifically chosen languages for the Creoles, we see that, although the source language (e.g. English or French) is most dominant, the other languages are still well distributed, with the exception of Yoruba. We surmise that the densely distributed, low-confidence scores for Yoruba can probably be attributed to the fact that Yoruba is a lower-resourced language.

2.6 EXPERIMENTS

In this section, we detail our experimental setups. We make our code and models publicly available.[5]

5 https://github.com/hclent/creole-dro

2.6.1 *Training*

Using the datasets described above, we conduct several experiments to assess how different training strategies affect the modeling of Creoles. We conduct all the experiments on both English BERT and multilingual mBERT models (Devlin et al., 2019). As our baseline, we consider pretrained $\text{BERT}_{\text{Base}}$ and mBERT models, and evaluate them on our development splits for the Creoles. We then assess the effectiveness of two popular training strategies: Empirical Risk Minimization (ERM) and Distributionally Robust Optimization (DRO). In this case, ERM consists of masked language modeling over all the data points in each dataset, in a similar fashion as done during pretraining.

For DRO, we utilize the WILDS library (Koh et al., 2020), which uses metadata associated with the input data to form the groups for DRO. In our case, we investigate three grouping strategies: grouping with language information as metadata (**DRO-Language**), as well as with two additional control experiments. In the first control experiment, we assign all training examples to the same group (**DRO-One**), such that that DRO is optimizing over only one large group. In the second control experiment, we randomly assign examples to one of four groups (**DRO-Random**). The motivation of for these control experiments is to ensure that improvements for DRO are actually grounded in the language information, and not an artifact of the WILDS grouping algorithm.

In **DRO-Language**, information about the examples' language makeup is used to determine the groups. In Mixed-Language, we rely on our knowledge of where the examples were sampled from, but in Creole-Only, we subdivide the Creole examples depending on their etymology. Specifically, grouping is done as follows in our two data setups outlined in sec:dro/ourdata:

- **Mixed-Language:** Here, grouping is done over the languages in the training data. For example, in the case of Nigerian Pidgin, if a sentence originally comes from the Yoruba corpus, it is assigned to the Yoruba group, and similarly for Nigerian Pidgin and the other languages listed in Table 2 for each Creole.

- **Creole-Only:** Here, as we only have the Creole samples, grouping is done over the confidence scores from the collection of the influential languages (see §2.5). An example is assigned to one of 2^N groups, representing the combinations of detected languages in a sentence. N is the number of languages listed in tab:split (Langs) for each Creole, and presence of a language is derived from its confidence score by the language identifier: if there is a confidence of 0.1% or higher that the language is represented in the sentence, then it is considered as present.

		Nigerian Pidgin			Singlish			Haitian Creole		
	BERT	P@1	P_D@1	PLL	P@1	P_D@1	PLL	P@1	P_D@1	PLL
	Pretrained	22.79	10.92	142.65	23.94	21.09	76.01	18.84	5.65	177.40
MIXED	ERM	**63.83**	**59.97**	**42.41**	**46.77**	**42.89**	**41.06**	**68.09**	**43.35**	**55.04**
	DRO-One	60.99	56.76	52.51	44.23	40.73	49.18	57.04	36.73	121.51
	DRO-Random	60.40	56.33	52.69	43.33	39.07	49.14	57.65	36.16	119.17
	DRO-Language	60.40	54.80	54.17	43.19	39.57	48.88	57.55	36.69	118.85
CREOLE-ONLY	ERM	**73.72**	**71.38**	**28.14**	**53.80**	**51.26**	**34.22**	**73.15**	**55.50**	**55.51**
	DRO-One	64.28	59.86	61.81	45.34	43.59	66.53	58.16	36.91	144.46
	DRO-Random	63.72	59.31	60.31	45.73	42.40	64.16	57.65	37.41	142.04
	DRO-Language	63.58	59.74	56.82	44.73	40.57	53.72	56.94	35.50	138.60

Table 3: Intrinsic evaluation: Precision@1 (P@1), Precision@1 for words in our Creole dictionary (P_D@1), and average Pseudo-log-likelihood score (PLL). We report results for MIXED-LANGUAGE (top) and CREOLE-ONLY (bottom). We note that ERM consistently outperforms the language models trained with robust objectives.

2.6.2 *Evaluation*

We perform two types of evaluation: intrinsic – based on the MLM training objective – and extrinsic – on traditional downstream NLP tasks.

INTRINSIC EVALUATION We evaluate our language models intrinsically with the following metrics:

- **Precision at k (P@k):** Precision of the language model in predicting a random masked token per sentence. This allows us to assess the general performance following the training objective. In the following, we report P@1. Results at $k = \{5, 10\}$ are in the App.
- **Dictionary-based precision at k (P_D@k):** Due to their nature, most of the words in a Creole sentence are from the corresponding source language (see fig:perm). Hence, for a more principled measurement of precision, we collect online dictionaries of our Creoles.[6] We perform the same MLM task as above, but this time only mask words belonging to the Creole dictionaries. By doing so, we can obtain a more accurate measure of what the

6 Nigerian Pidgin: http://naijalingo.com/.
Singlish: http://www.mysmu.edu/faculty/jacklee/.
Haitian Creole: https://kreyol.com/dictionary.html.

LMs have learned. We again report results at $k = 1$ here, and refer the reader to the App. for $k = \{5, 10\}$.

- **Mean pseudo-log-likelihood score (PLL):** Following recent studies (Salazar et al., 2020; Shin, Lee, and Jung, 2019; Wang and Cho, 2019), we measure the pseudo-log-likelihood scores from MLMs given by summing the conditional log probabilities $\log \mathbb{P}_{\text{MLM}}(w_t|\mathbf{w}_{\setminus t})$ of each token w_t in a sentence $\mathbf{w} = \langle w_1, \ldots, w_T \rangle$. These are obtained in BERT by replacing w_t with the special `[MASK]` token. Here, we report the mean score given by:

$$\text{PLL} = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{w} \in \mathcal{C}} \frac{1}{|\mathbf{w}|} \sum_{w_t \in \mathbf{w}} \log \mathbb{P}_{\text{MLM}}(w_t|\mathbf{w}_{\setminus t}; \theta), \tag{1}$$

where $\mathcal{C}$ denotes the evaluation corpus, and θ denotes a model's parameters.

EXTRINSIC EVALUATION We also perform an extrinsic evaluation of our models on downstream tasks, for the datasets that are available. Specifically, we train and evaluate models for Nigerian Pidgin NER and POS tagging with Universal Dependencies (Nivre et al., 2020a, UPOS), as well as Singlish UPOS. We fine-tune our pretrained language models on the training sets of these two tasks and evaluate them on the corresponding test sets.

2.6.3 *Framework*

We write our code in PyTorch (Paszke et al., 2019). In particular, for language model training, we rely on the HuggingFace Transformers library (Wolf et al., 2019), and the WILDS library (Koh et al., 2020) for DRO. Models are fine-tuned for 100,000 steps with batch size of 16. For downstream tasks, we use MaChAmp (Goot et al., 2021a) and train our models for 10 epochs. The best checkpoints were selected based on performance on the dev sets. Unless otherwise specified, we use the default hyperparameters. Our experiments are run on one NVIDIA TitanX GPU in a shared cluster.

2.7 RESULTS AND ANALYSES

INTRINSIC EVALUATION The main finding of the intrinsic evaluation is that ERM outperforms DRO for all grouping strategies across all metrics. We also observe that P_D@k is a more difficult task than the standard precision at k, with randomly masked tokens (see A.1 for full results with both BERT and mBERT). Moreover we find that the DRO models often have a much higher perplexity than ERM. Finally, the results show that, between the MIXED-LANGUAGE and CREOLE-ONLY experiments, the latter performed better, demonstrating that

		Nigerian Pidgin		Singlish
	BERT	NER [F_1]	UPOS [Acc]	UPOS [Acc]
MIXED	ERM	87.86	98.00	**91.24**
	DRO-Language	**88.40**	**98.06**	90.22
C-ONLY	ERM	**87.98**	**98.04**	**91.17**
	DRO-Language	87.12	97.98	90.44

Table 4: Extrinsic evaluation. Similar performance on downstream tasks across all models demonstrate show that language model training did *not* benefit significantly from neither DRO nor data in related languages.

training on additional data was not useful for learning language models for Creoles. While we only report results for BERT here, we observe the same patters with mBERT (see A.1).

EXTRINSIC EVALUATION Here, we observe the same trend as in the intrinsic evaluation: ERM performs better than DRO (see Table 4). Although for Nigerian Pidgin DRO-Language performs better than ERM on both NER and UPOS, the gap between the scores is too small to draw concrete conclusioins from.

There are several factors that could have influenced the DRO models to perform worse than ERM. We explore their effects below.

OVER-PARAMETERIZATION Over-parameterization is known to be problematic for DRO (Sagawa et al., 2019). In order to investigate the role of over-parameterization in our experiments, we ran additional MIXED-LANGUAGE experiments on Nigerian Pidgin English, with different sized BERT models, namely $\text{BERT}_{\text{Tiny}}$, $\text{BERT}_{\text{Small}}$ (Jiao et al., 2020), and $\text{BERT}_{\text{Base}}$. The results in Table 5 demonstrate that over-parameterization was not a leading cause for DRO failure, otherwise we would expect for smaller BERT versions to have relative better performance compared to the corresponding ERM runs. Instead, we see that standard BERT works fine for this task, and over-parameterization is not the cause of poor performance of DRO in our experiments.

REGULARIZATION Sagawa et al. (2019) also discuss how lack of regularization lead to problems for DRO, and how increased regularization is necessary for worst-group generalization. To investigate this potential weakness in our experiments, we run additional experiments using $\text{BERT}_{\text{Small}}$ on MIXED-LANGUAGE data for Nigerian Pidgin English, trying different weight decay values in each Table 6. If our DRO models were suffering from insufficient regularization, we

		Nigerian Pidgin		
BERT	**Size**	P@1	P_D@1	PLL
ERM	Tiny	31.31	26.12	110.23
	Small	47.39	46.75	77.47
	Base	63.83	59.97	42.41
DRO-Language	Tiny	31.00	23.09	99.70
	Small	43.00	37.75	82.50
	Base	60.40	54.80	54.17

Table 5: Over-parameterization experiments with MIXED-LANGUAGE Nigerian Pidgin English data. Smaller sized models do not benefit DRO over ERM.

		Nigerian Pidgin		
BERT	**Weight Decay**	P@1	P_D@1	PLL
ERM	0.01	47.39	46.75	77.47
DRO-Language	0.01	43.00	37.75	82.50
	0.05	42.86	38.47	83.03
	0.10	43.00	38.74	81.80
	0.30	42.70	39.53	81.94

Table 6: Regularization experiments on MIXED-LANGUAGE Nigerian Pidgin data, based on $\text{BERT}_{\text{Small}}$.

would expect that increasing the regularization factor of weight decay would boost performance. However, we find no meaningful effect of this hyperparameter, which leads us to believe that insufficient regularization is not a driving factor in the underperformance of DRO compared to ERM.

DRIFT AND CREOLE STABILITY Creole languages arise from pidgins, which are initially developed for use as second language. Recent years have seen renewed interest in the classic question of the relationship between pidgin and Creole formation and second language acquisition (Plag, 2009). To investigate the matter of Creole stability, we follow (Ben-David et al., 2007) and calculate the proxy $\mathcal{A}$-distance (PAD) between different domains of Creole data (see Table 7). Specifically, we train an SVM on the BERT encodings.[7] Our $\mathcal{A}$-distance results suggest that Creole languages do *not* exhibit more drift than English when the data are comparable. This potentially explains why

7 Our code is adapted from https://github.com/rpryzant/proxy-a-distance.

Language	Domain-1	Domain-2	PAD
English	Disaster Response Corpus	Newswire	1.75
Haitian Creole	Disaster Response Corpus	Newswire	1.47
English	EWT-UD	NUD	1.04
Nigerian	UNMT	NUD	1.28

Table 7: Proxy $\mathcal{A}$-distance (PAD) scores on parallel (Haitian) or near-parallel (Nigerian) data. PAD is proportional to domain classification error; hence, large distances mean high domain divergence. Our results suggest that Creole languages do *not* exhibit significantly more drift than other languages.

distributionally robust language models do not outperform regular language models trained with empirical risk minimization objectives.

2.8 CONCLUSION

In this paper, we bring Creole languages to the attention of the NLP community. We collect data and train baseline language models for three Creoles, and evaluate these models across the downstream tasks of part-of-speech tagging and named entity recognition. Based on previous work suggesting the instability of Creole languages (Patrick, 1999; Winford, 1999), we explore the impact of using more robust learning objectives for masked language modeling of Creoles, but our results show that vanilla empirical risk minimiziation is superior. We show that this is not the result of over-parameterization or lack of regularization, but instead suggest this is a result of the relative stability of Creole languages. We note that it still remains possible that significant improvements could be achieved by modeling dynamics specific to Creole languages, i.e., the processes that govern their development, including social factors (Holm, 2000) and second language acquisition dynamics (Plag, 2009).

3

ANCESTOR-TO-CREOLE TRANSFER IS NOT A WALK IN THE PARK

3.1 ABSTRACT

We aim to learn language models for Creole languages for which large volumes of data are not readily available, and therefore explore the potential transfer from ancestor languages (the 'Ancestry Transfer Hypothesis'). We find that standard transfer methods do not facilitate ancestry transfer. Surprisingly, different from other non-Creole languages, a very distinct two-phase pattern emerges for Creoles: As our training losses plateau, and language models begin to overfit on their source languages, perplexity on the Creoles *drop*. We explore if this *compression* phase can lead to practically useful language models (the 'Ancestry Bottleneck Hypothesis'), but also falsify this. Moreover, we show that Creoles even exhibit this two-phase pattern even when training on random, unrelated languages. Thus Creoles seem to be typological outliers and we speculate whether there is a link between the two observations.

3.2 INTRODUCTION

Creole languages refer to vernacular languages, many of which developed in colonial plantation settlements in the 17th and 18th centuries. Creoles most often emerged as a result of contact between social groups that spoke mutually unintelligible languages, i.e., from the interactions of speakers of nonstandard varieties of European languages and speakers of non-European languages (Lent et al., 2021a). Some argue these languages have an exceptional status among the world's languages (McWhorter, 1998), while others counter that Creoles are not unique, and evolve in the typical manner as other languages (Aboh and DeGraff, 2016). In this paper, we will present experiments in evaluating language models trained on non-Creole languages for Creoles, as well as in various control settings. We first explore the following hypothesis:

R1: Language models trained on ancestor languages should transfer well to Creole languages.

We call **R1** the 'Ancestry Transfer Hypothesis.' Our experiments, however, suggest that **R1** is *not* easily validated. We note, though, that ancestor-to-Creole training exhibits divergent behavior when training *for long*, leading to the following hypothesis:

R2: Language models trained on ancestor languages can, after a compression phase, transfer well to Creole languages.

We call **R2** the 'Ancestry Bottleneck Hypothesis.' While compression benefits transfer, performance never seems to reach useful levels. Furthermore, similar effects are observed with Creoles when training on non-ancestor languages. Our findings here are not relevant to applied NLP, but they shed light on cross-lingual training dynamics (Deshpande, Talukdar, and Narasimhan, 2021; Singh et al., 2019), and we believe they have potential implications for the linguistic study of Creoles (DeGraff, 2005c), as well as for information bottleneck theory (Tishby, Pereira, and Bialek, 1999).

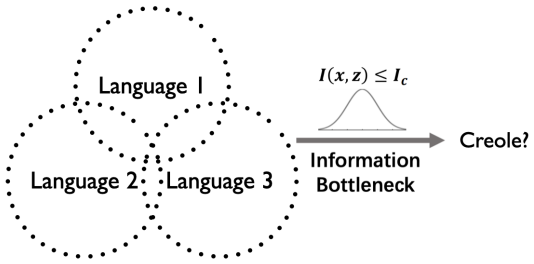


Figure 4: Does the Information Bottleneck principle capture some of the dynamics of Creole formation?

OUR CONTRIBUTIONS We conduct a large set of experiments on cross-lingual zero-shot applications of language models to Creoles, primarily to test whether ancestor languages provide useful training data for Creoles (the 'Ancestry Transfer Hypothesis;' **R1**). Our results are a mix of negative and positive results: **First Negative Result:** Ordinary transfer methods do not enable ancestor-to-Creole transfer. **First Positive Result:** Regardless of the source languages, when training for long periods of time, a compression phase takes places for Creoles: as the models overfit their training data, perplexity on Creoles begin to decrease. This pattern is unique to Creoles as it does not emerge for target non-Creole languages. **Second Negative Result:** The compression phase does not lead to better representations for downstream tasks in the target Creoles.

3.3 BACKGROUND

CROSS-LINGUAL TRAINING DYNAMICS Several multilingual language models have been presented and evaluated in recent years. Since Singh et al. (2019) showed that mBERT (Devlin et al., 2019) generalizes well across related languages, but compartmentalizes language families, several researchers have explored the training dynamics of training multilingual language models across related or distant

Creole	Ancestors	Random Controls
pcm	eng, hau, yor, por	afr, chr, hun, quy
jam	eng, hau, spa, ibo	afr, chr, hun, quy
acf	fra, hau, spa, ibo	afr, chr, hun, quy
hat	fra, fon, spa, ibo	afr, chr, hun, quy
Non-Creole	**Relatives**	**Random Controls**
spa	fra, ita, por, rom	afr, chr, hun, quy
dan	nno, isl, swe, deu	afr, chr, hun, quy

Table 8: Transfer setups in our study. We aim to learn target Creoles and Non-Creoles by training on **1)** their Ancestors or Relatives, respectively; and **2)** languages unrelated to the target ones as a control (Random Controls).

language sets (Deshpande, Talukdar, and Narasimhan, 2021; Keung et al., 2020; Lauscher et al., 2020). Unlike most previous work on cross-lingual training, we focus on evaluation on unseen (Creole) languages. This set-up is also explored in previous work focusing on generalization to unseen scripts (Muller et al., 2021; Pfeiffer et al., 2021). Muller et al. (2021) argue that generalization to unseen languages is possible for seen scripts, but hard or impossible for unseen scripts, but this paper identifies a third category of unseen languages with seen scripts, which exhibit non-traditional learning curves in the zero-shot pre-training regime.

LINGUISTIC THEORIES OF CREOLE Creolists have long debated whether Creole languages have an exceptional status among the world's languages (DeGraff, 2005a). McWhorter (1998) argues that Creoles are *simpler* than other languages, and defined by minimal usage of inflectional morphology, little or no use of tone encoding lexical or syntactic contrasts, and generally semantically transparent derivation. Others have argued that Creoles cannot be be unambiguously distinguished from non-Creoles on strictly structural, synchronic grounds (DeGraff, 2005a). On this view Creole grammars do not form a separate typological class, but exhibit many similarities with the grammars of their parent languages, e.g., the similarities in lexical case morphology between French and Haitian Creole. We do not take sides in this debate, but observe that the exceptionalist position would explain our results that zero-shot transfer to Creole languages is particularly difficult. Exceptionalism also aligns well with the heatmaps presented in §3.6.

INFORMATION BOTTLENECK The Information Bottleneck principle (Tishby, Pereira, and Bialek, 1999) is an information-theoretic framework for extracting output-relevant representations of inputs, i.e., com-

pressed, non-parametric and model-independent representations that are as informative as possible about the output. Compression is formalized by mutual information with input. A Lagrange multiplier controls the trade-off between these two quantities (informativity and compression). Being able to compute this trade-off assumes the joint input–output distribution is accessible. The trade-off is found by ignoring task-irrelevant factors and learning an invariant representation. The intuition behind the 'Ancestry Bottleneck Hypothesis' (**R2**) is that invariant representations are particularly useful for Creoles (see Figure 4 for an illustration).

3.4 MULTILINGUAL TRAINING

This section sets out to evaluate the 'Ancestry Transfer Hypothesis' (**R1**). To this end, we evaluate multilingual language models – trained with a BERT architecture from scratch, but of smaller size and with less data (Dufter and Schütze, 2020) – on Creoles such as Nigerian Pidgin or Haitian Creole. We compare two scenarios: **1)** a scenario in which the training languages are languages that are said to be *parent* or *ancestor* languages of the Creole, such as French to Haitian, and **2)** a scenario in which *random*, unrelated training languages were selected. To compare against Creoles, we also explore these transfer scenarios for two target non-Creoles – Spanish and Danish – training on languages closely related to them (i.e., as typically done in cross-lingual learning). Table 8 lists all the transfer scenarios that we investigated. Our experimental protocol follows Dufter and Schütze (2020), and it is described in detail below.

We aim to learn language models for Creole languages for which large volumes of data are not readily available, and therefore explore the potential transfer from ancestor languages (the 'Ancestry Transfer Hypothesis'). We find that standard transfer methods do not facilitate ancestry transfer. Surprisingly, different from other non-Creole languages, a very distinct two-phase pattern emerges for Creoles: As our training losses plateau, and language models begin to overfit on their source languages, perplexity on the Creoles *drop*. We explore if this *compression* phase can lead to practically useful language models (the 'Ancestry Bottleneck Hypothesis'), but also falsify this. Moreover, we show that Creoles even exhibit this two-phase pattern even when training on random, unrelated languages. Thus Creoles seem to be typological outliers and we speculate whether there is a link between the two observations.

EXPERIMENTAL PROTOCOL We train BERT-smaller models (Dufter, Schmitt, and Schütze, 2020), consisting of a single attention head (shown to be sufficient for achieving multilinguality by K et al. 2020). Although training smaller models means our results are not directly

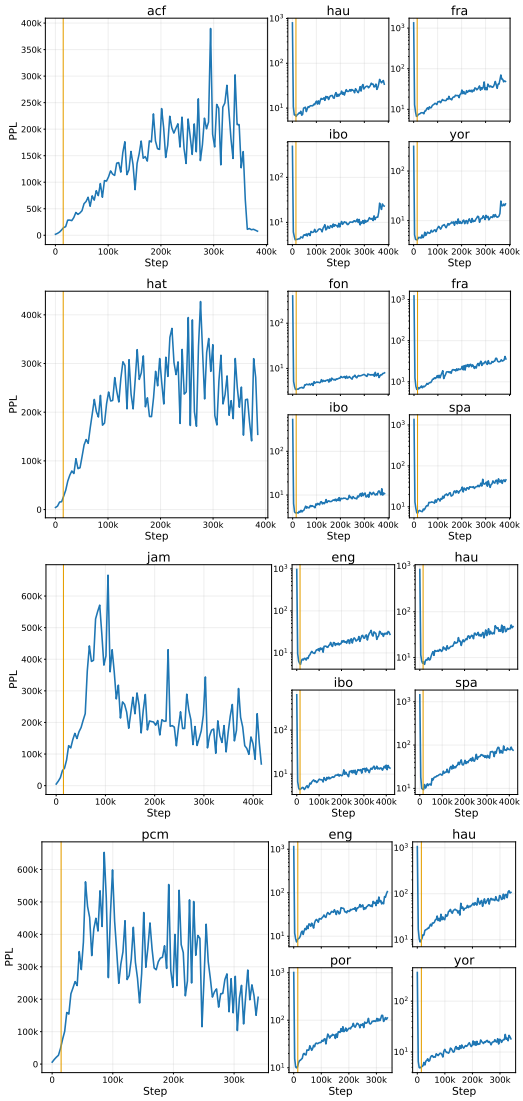


Figure 5: Four zero-shot transfer experiments for Creole languages. The left-hand side plot shows the (zero-shot) validation curve for checkpoints on Creole data; the small plots show the learning curves for the training languages. We see an initial increase in perplexity (disproving **R1**). The yellow vertical line denotes 100 epochs. We also see a subsequent decrease in perplexity.

comparable to larger models like mBERT or XLM-R (Conneau et al., 2019), there is evidence to support that smaller transformers can work better for smaller datasets (Susanto, Htun, and Tan, 2019), and that the typical transformer architecture would likely be overparameterized for our small data (Kaplan et al., 2020). Thus, the BERT-smaller models appear to be the most appropriate match for our very small datasets. The models are trained on a multilingual dataset, consisting of an equal parts of each source language, taken from the Bible Corpus (Mayer and Cysouw, 2014). We chose Bible data to train our models as it facilitates a controlled setup with parallel data in many languages whilst including our low-resource Creoles and ancestors. For

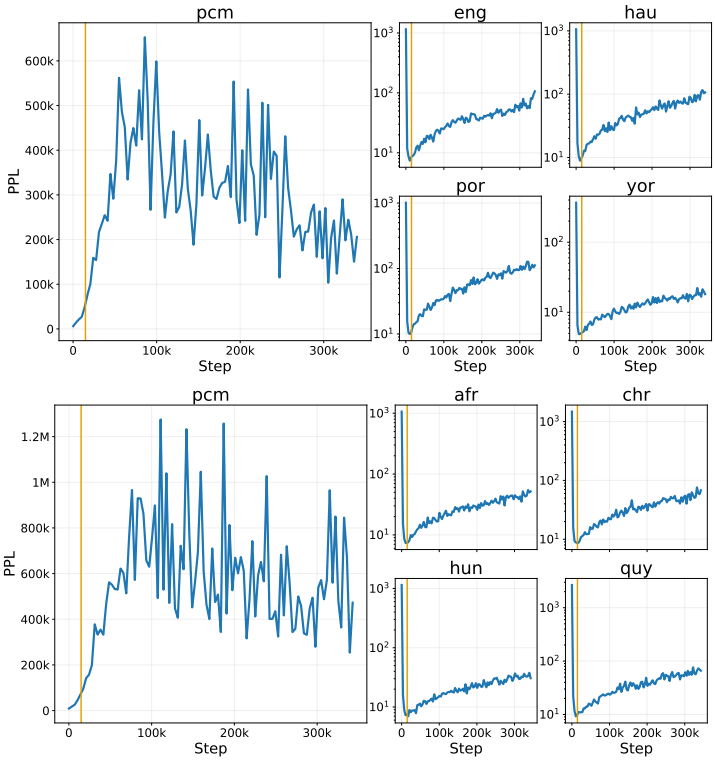


Figure 6: Learning curves for Nigerian Pidgin English when training on **ancestor** languages (top) and when training on **random** languages (bottom). No significant differences are observed. This disproves **R2**.

each experiment, we learn a custom BERT tokenizer on source and target languages, with a vocabulary size of 10,240 word pieces (Wu et al., 2016).[1] Each model is trained for 100 epochs (see Table 9).

We also follow Dufter and Schütze (2020)'s approach of calculating the perplexity on 15% of randomly masked tokens (w), with probabilities (p), as $\exp(-1/n \sum_{k=1}^{n} \log(p_{w_k}))$. We calculate perplexity on held out development data for both source and target languages. Our code is available online.[2]

RESULTS In Figure 5, by 100 epochs (indicated by a yellow vertical line), we observe two different patterns for Creoles and non-Creoles. For target Creole languages, the models are able to learn the ancestor languages, but perplexity on the held out Creoles consistently climbs. On the other hand, for target non-Creoles, we observe a slight initial drop in perplexity before it starts to increase as the models overfit the source languages.

1 We explored different vocabulary sizes (1,024, 2,048 and 10,240) as well as other tokenization techniques (grapheme-to-phoneme and byte-pair encodings Sennrich, Haddow, and Birch 2016), which did not affect the overall findings discussed below.

2 https://github.com/hclent/ancestor-to-creole

Hyperparameter	Creole	Non-Creole
Vocabulary size	10,240	10,240
Learning rate	1.00E-04	5.00E-05
Weight decay	1.00E-03	1.00E-03
Dropout	1.00E-01	1.00E-01
Batch size	256	256

Table 9: The hyperparameters used for target Creole and Non-Creole experiments. Vocab size, weight decay, and dropout were the same across Creole and Non-Creole experiments, however the Non-Creoles required a smaller learning rate, in order to successfully learn. All experiments were run on a TitanRTX GPU.

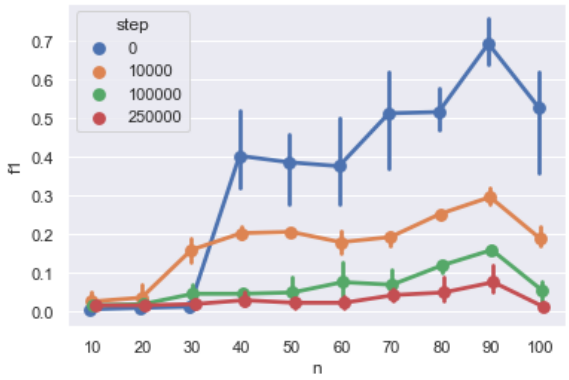


Figure 7: Results for downstream performance on Nigerian Pidgin NER, across 3 random seeds. The top row shows our model trained on ancestor of Nigerian Pidgin (pcm), while the bottom one shows results for mBERT. Step 0 in the legend refers to the pre-trained mBERT, without any further training on ancestor languages.

3.5 TRAINING FOR LONGER

It seems linguistically plausible that training for longer on ancestor languages to learn more invariant representations should better facilitate zero-shot transfer to Creole languages. This is the essence of the 'Ancestry Bottleneck Hypothesis' (**R2**), which we explore in this section.

CREOLE COMPRESSION We continue training our models for 5 days, for each Creole and non-Creole target language – which typically results in 300k–500k steps of training (and thus, extremely overfit). As the models overfit to the source languages, we observe a notable drop in perplexity for Creoles, which is true *regardless* of the training data (ancestors versus random controls), as shown in Figure 5 and Figure 6. On the other hand, these plots show that this compression does not emerge for non-Creole target languages, as their complexity steadily increases as the models overfit their training data more and more.

DOWNSTREAM PERFORMANCE Next, in order to determine if this compression present for Creoles can be beneficial, we used MACHAMP (Goot et al., 2021b) to check the ability of our Nigerian Pidgin models to fine-tune for downstream NER (Adelani et al., 2021). We evaluate the representations learned at different stages of pre-training by fine-tuning our checkpoints corresponding to early stage (10,000 steps), maximum perplexity, and post-compression (last checkpoint). Each model is fine-tuned for 10 epochs. Figure 7 shows that, across three random seeds, post-compression checkpoints consistently perform worse than pre-compression or max-complexity checkpoints. The results negate **R2**, i.e., that the compression effect observed during training would be useful for Creoles.[3]

FEW-SHOT LEARNING Finally, we assess the ability of our models to learn Creoles from few examples (n=10, ..., 100) at different training stages. Once again, few-shot learning from post-compression checkpoints led to higher perplexity than training from maximum perplexity or early checkpoints.

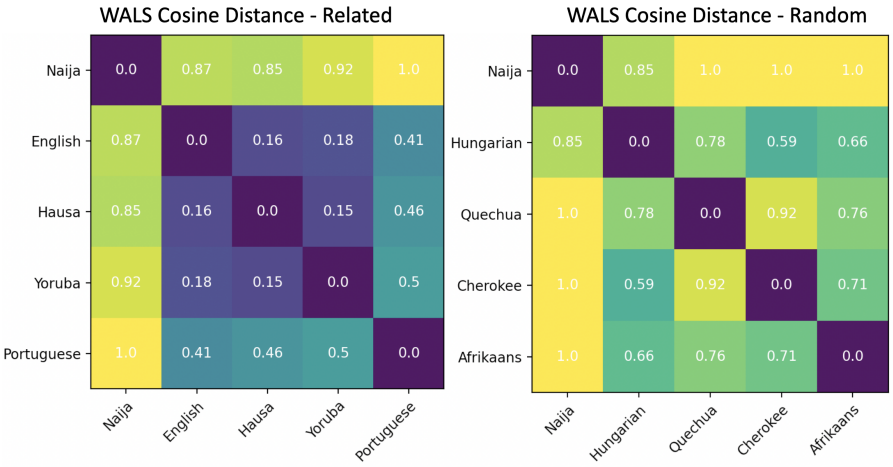


Figure 8: Heatmaps of WALS cosine distances between Nigerian Pidgin (Naija) and its parent and random training languages. We observe that Nigerian Pidgin is *less* related to any of these languages, than any of them internally (except Quechua and Cherokee).

3.6 CREOLES THROUGH THE LENS OF WALS

We have observed unique patterns for Creoles. Namely, multilingual learning of the related languages did not lead to successful transfer to Creoles; and that Creoles exhibit a unique compression effect. Here, we speculate whether there is a link between these observations, and investigate whether typological features can shed lights into our

3 We also compared the results of a pre-trained mBERT, which, unsurprisingly, outperformed all of our checkpoints (corresponding to smaller models learned from tiny data).

results. To that effect, we use The World Atlas of Language Structures (WALS)[4], which has been used to study Creoles before (Daval-Markussen and Bakker, 2012). Here, we use the cosine distance between the normalized (full) WALS feature vectors as our distance metric.[5]

In Figure 8, we present an example heatmap for Nigerian Pidgin, which shows that Nigerian Pidgin is *less* related to ancestor and random languages than any of them internally (except Quechua and Cherokee). We found this pattern present for each of the Creoles. Thus, it would seem that Creoles' relatively large distance[6] from other languages may make cross-lingual transfer a particular challenge for learning Creoles.[7]

3.7 CONCLUSION

We have presented two hypotheses (**R1** and **R2**) about the possibility of zero-shot transfer to Creoles, both built on the idea that Creoles share characteristics with their ancestor languages. This is not exactly equivalent to the so-called superstratist view of Creole genesis, which maintains that Creoles are essentially regional varieties of their European ancestor languages, but if the superstratist view was correct, **R1** would very likely be easily validated (Singh et al., 2019). Our results show the opposite trend, however. Zero-shot transfer to Creole languages from their ancestor languages is hard. We do not claim that our results favor an exceptionalist position on Creoles. While we performed a first analysis of several segmentation approaches (i.e., BERT word piece, grapheme-to-phoneme, and byte-pair encodings) – which did not change the training dynamics – we believe that a rigorous comparison would be beneficial for future work in ancestor-to-Creole transfer. We hope that continued investigation in this direction can shed more light on cross-lingual transfer, especially with regards to Creoles, and that this work has demonstrated that not all transfer between related languages is trivial.

4 wals.info.

5 https://github.com/mayhewsw/wals.

6 We note that previous work has suggested that WALS features alone may be insufficient for typological comparison of Creoles to non-Creoles (Murawaki, 2016).

7 We also note that cosine distance might not be meaningful here, as the normalized (full) space does not represent the feature geometry of the space that the linguists that developed the features in WALS were assuming.

4

WHAT A CREOLE WANTS, WHAT A CREOLE NEEDS

4.1 ABSTRACT

In recent years, the natural language processing (NLP) community has given increased attention to the disparity of efforts directed towards high-resource languages over low-resource ones. Efforts to remedy this delta often begin with translations of existing English datasets into other languages. However, this approach ignores that different language communities have different needs. We consider a group of low-resource languages, *Creole* languages. Creoles are both largely absent from the NLP literature, and also often ignored by society at large due to stigma, despite these languages having sizable and vibrant communities. We demonstrate, through conversations with Creole experts and surveys of Creole-speaking communities, how the things needed from language technology can change dramatically from one language to another, even when the languages are considered to be very similar to each other, as with Creoles. We discuss the prominent themes arising from these conversations, and ultimately demonstrate that useful language technology cannot be built without involving the relevant community.

4.2 INTRODUCTION

The field of natural language processing (NLP) has become aware that most of the world's languages are unfortunately under-represented, or entirely absent, from the field's body of work (Joshi et al., 2020a). In recent years, there has been a push in efforts to ameliorate this discrepancy (Mirzakhalov et al., 2021; Nekoto et al., 2020; Ogueji, Zhu, and Lin, 2021). Among these low-resourced languages[1] are *Creole* languages, which are particularly under-resourced due to barriers like societal stigma (Siegel, 1999), despite the fact that these languages are spoken by many people globally. One line of work has focused on creating datasets for low-resource languages via the translation of existing high-resource language datasets (Artetxe, Ruder, and Yogatama, 2020; Budur et al., 2020; Conneau et al., 2018). Despite the popularity of this method, it poses several issues, which can negatively affect the communities of these low-resource languages. One such issue lies in translation artifacts, which have been shown to have notable impacts

1 This term is often largely ambiguous, and all "low-resource" languages should not be conflated together into one large group, but rather considered independently, in the context of its speakers, their culture, and their needs.

on the performance of models trained with such datasets (Artetxe, Labaka, and Agirre, 2020). Furthermore, translated datasets are often simplified and unnatural, a phenomenon referred to as *translationese* (Volansky, Ordan, and Wintner, 2013). This has also been shown to adversely affect the evaluation of machine translation models (Graham, Haddow, and Koehn, 2020). Creoles, too, are not immune to the shortcomings of this approach. Moreover, many translated datasets will simply not be relevant to communities speaking a Creole language, as concepts relevant to the original high-resource source language are subsequently translated into the low-resource language, despite being irrelevant to people or cultures speaking the language (Liu et al., 2021a). For example, sentences about American football or the American Thanksgiving holiday are simply not relevant or necessary for speakers of Creole languages. The same mismatch also applies to other more geographical-specific domain information present in the data, such as landmarks or landscapes. All of these show that, while there may be good intentions behind this approach, it could potentially lead to poor models for speakers of low-resource languages and even to the creation of tools of little use or relevance for Creole speakers.

Meanwhile, works such as Hu et al. (2011c) concretely demonstrate how crowd-sourcing data from target-language speakers, even if monolingual, leads to improved results for statistical machine translation systems. While these findings are not up to date with contemporary neural machine translation, involving native speakers minimizes the risk of having non-relevant examples included in the dataset. However, as the authors also note, there can be considerable logistical difficulties of finding native speakers to contribute, even when offering payment. And even if one manages to recruit paid speakers, a large problem still remains: the underlying exploitative nature of treating language speaking communities like data resources to be mined. Bird (2020) discusses in detail these foundational problems within the language technology community, and how, in order to break the cycle of harmful colonialism in our science, we must fundamentally change the relationship between researchers and the language-speaking communities. But the only way we can learn this, claims Bird (2020), is by establishing a respectful, “feedback/collaboration loop", and necessarily involving community members in our research.

Following the work of Bird (2020), in this work, we focus on the problem of creating resources for low-resource languages, in this case Creoles, and the inherent presupposition by researchers of what technologies are indeed wanted and needed by the communities speaking those languages. While many researchers may assume that the “best-case scenario" for all languages would be to have all language technologies *equally* available, the fact of the matter is that many com-

munities have very specific wants and needs of language technology, as well as language technologies that are notably unwelcome, even though they are a commonplace for high-resource languages. Disregarding the needs of a language community can lead to misuse of finite resources on creating unnecessary datasets or technologies while leaving the community's highest priorities neglected. And finally, when researchers assume what technologies are wanted on the behalf of a community, it inherently alienates that community and takes away their agency (Bird, 2020). In this work, we explore how the needs of different Creole-speaking communities vary wildly from one another, and we demonstrate the need to establish respectful relationships with experts and communities, in order to make truly useful language technology.

Our contributions in this work are as follows:

- We present a survey of Creole NLP, and discussion of features from Creole languages that present unique challenges to existing NLP workflows.
- We discuss important considerations, gleaned from conversations with experts, and a survey of Creole language speakers.
- We propose a Creole continuum for language technology, as a guiding framework of research considerations, to help NLP researchers planning to work on Creoles.

4.3 BACKGROUND

Today, Creole languages are spoken widely throughout the Caribbean and West Africa, as well as parts of South America, Asia, Australia and the Pacific. Creoles have long captured the attention of linguists due to their unique, and sometimes tragic[2], histories with regards to language evolution. Typically, Creole languages originate from situations in which multiple different languages have come into close contact with each other (Thomason and Kaufman, 1992). The exact process of how a Creole language is "born" (i.e. Creole genesis), as well as discussion of which linguistic features a Creole inherited from the various "parent" languages, have been the subject of intense and ongoing linguistic debate for decades (Alleyne, 1971; Bickerton, 1984; Muysken and Smith, 1986; Sessarego, 2020). On one hand, some believe that Creoles themselves form a unique typological class of languages, with a separate place on the phylogenetic tree of languages (i.e. Creole exceptionalism, Bickerton (1984)). Linguists supporting Creole exceptionalism typically claim that Creoles are more simple than other languages (Parkvall et al., 2008), for example, lacking in

2 For example, Caribbean Creoles resulted from the displacement of African peoples in the Atlantic slave trade.

complex morphology (McWhorter, 1998). On the other hand, others argue that there are no grounds to claim that Creole evolution is especially different from the language evolution of so-called "normal" (DeGraff, 2003, 2005b). And indeed, Creoles do exhibit behaviors just as complex as non-Creole languages (DeGraff, 2001), including complex morphology (Henri, Stump, and Tribout, 2020).

Moreover, some criticisms of Creole exceptionalism also examine how the history of Creole studies itself has unfortunately been riddled with discrimination and racism (DeGraff, 2005b). In the past, Creoles were often considered to be something short of a full-fledged language (or, more harshly, "degenerate variants or dialects of their parent languages"[3]). According to Kouwenberg and Singler (2009), "A part of the legacy of slavery in the Caribbean and elsewhere has been the stigmatization of the languages associated with slaves ... [the] willingness to apply the concept of linguistic relativism – whereby every language is understood to be complete and valid – may have been extended to Hopi and Hausa[4], but it generally stopped short of being extended to Creoles." In line with this, in this work, we hope to raise awareness in the NLP community about why Creoles are important to work with. Beyond being the subject of vibrant linguistic debate, Creoles are often ignored when it comes to language technology, which puts speakers of already often stigmatized languages at a further disadvantage. For the remainder of this section, we will present a survey of existing Creole datasets, a summary of works published on NLP for Creoles, and finally end this section with a discussion of some specific features of Creole languages that are notable within the context of NLP.

4.3.1 *Creole Data and Creole NLP*

In this section we will detail existing resources and datasets for Creole languages (including those which are now seemingly defunct), as well as discuss related works actively focused on NLP for Creoles.

VERIFIED RESOURCES Although Creole languages are in general very low-resourced, the datasets that do exist vary widely from task to task, as well as from language to language. Hagemeijer et al. (2014) presents an extensive overview of Creole data resources through 2014 for a wide variety of Creoles, many of which are more traditional corpora, (e.g., transcriptions of conversations made by linguists with formal training, or scans of documents originally written in the Creole language); though these may not have the relevant annotations for common NLP tasks. Lent et al. (2021a) also provides a thorough

3 https://en.wikipedia.org/wiki/Creole_language#Overview

4 Hopi is an Native American indigenous language from Arizona, United States; Hausa is a Chadic language, spoken in West and Central Africa.

Language	Resource	Description	Status
Haitian Kreyol	Haitian Disaster Response Corpus (Callison-Burch et al., 2011; Munro, 2010)	SMS	Verified
	CMU Haitian Corpus http://www.speech.cs.cmu.edu/haitian/	Speech and Text Corpora	Verified
Hawaiian Pidgin	Multilingual Hawai'i Linguistic Landscape Corpus (Purschke, 2021)	Image Repo with Annotations	Verified
Reunionese Creole & Seychellois Creole	Creolica http://creolica.net/	Text and Short Stories in HTML or PDFs	Verified
Singlish	National University of Singapore SMS Corpus (Chen and Min-Yen, 2015b)	SMS	Verified
	Universal Dependencies for Colloquial Singaporean English (Wang et al., 2017)	UD Treebank	Verified
	Webcrawler for Singaporean Hardware Forum (Tan et al., 2020)	Webcrawler	Verified
Sri Lankan Malay (Endangered)	The Language Archive dobes	Audio and XML	Verified

Table 10: Descriptions of every Creole resource or dataset that we could identify and also verify as being readily available online. (Part 1/2)

overview of existing NLP datasets for Haitian Kreyol, Singaporean Colloquial English (Singlish), and Nigerian Pidgin English. In this work, we set about the task of manually verifying each dataset presented by Hagemeijer et al. (2014) and Lent et al. (2021a), as well as searching for additional resources. We present all "verified" datasets in Tables 10 and 11. Here, we use "verified" to mean that we could easily find the resource described in the paper, through either a provided URL in a publication, or through a search engine.

Readers should note that we excluded both extinct Creoles and ostensibly historical Creole data from Tables 10 and 11. Those interested can see that there are available data for the extinct Virgin Islands Dutch Creole.[5] Other historical Creole data include the Corpus of Mauritian Creole Texts (Baker and Sing, 2007), a collection of texts spanning the 1730 to 1930, and the Surinam Creole Archive (suca.ruhosting.nl), which should have historical texts for both Sranan

5 doecreoltaal.com

Language	Resource	Description	Status
Nigerian Pidgin	NaijaSynCor (Bigi, Caron, and Abiola, 2017)	Speech Recognition	Verified
	JW300 Corpus (Agić and Vulić, 2019)	Parallel Texts for Machine Translation	Verified
	Pidgin UNMT (Ogueji and Ahia, 2019)	Monolingual Texts for Machine Translation	Verified
	Naija-English Codeswitching Corpus (Ndubuisi-Obi, Ghosh, and Jurgens, 2019b)	News Articles with Comments; Annotated for code switching	Verified
	Surface-Syntactic UD Treebank for Naija (Caron et al., 2019)	Universal Dependencies	Verified
	Speech-to-Text Nigerian Pidgin Dataset (Ajisafe et al., 2020)	Speech Recognition	Verified
	NaijaNER (Oyewusi et al., 2021a)	Named Entity Recognition	Verified
	Masakhaner (Adelani et al., 2021)	Named Entity Recognition	Verified
	NaijaSenti (Muhammad et al., 2022)	Sentiment Analysis	Verified

Table 11: Descriptions of every Creole resource or dataset that we could identify and also verify as being readily available online. (Part 2/2)

Tongo and Saramaccan, although the hyperlinks are presently broken in this website.

Lastly, to utilize linguistic information about Creoles, the Atlas of Pidgin and Creole Language Structures (APiCS) is an indispensable resource (Michaelis et al., 2013). APiCS is an extension of the popular WALS resource (Dryer and Haspelmath, 2013), but is solely dedicated to pidgins and Creoles.

UNVERIFIED RESOURCES Unfortunately many of the Creole corpora reviewed in Hagemeijer et al. (2014) are no longer available, with broken URLs. We describe any resource as “not verifiable", when we cannot track down the resource through the combination of a URL, a simple web search, or through the original publication. These resources may still exist, but they are too difficult to find with a reasonable effort made. The list of “not verifiable" resources can be found in Tables 12 and 13. We hope that highlighting the “not verifiable" datasets can serve as a call to action in the field, to consider long term data hosting solutions. In order to make the information we gathered about datasets useful in the long-term, we release a community-based

Language	Resource	Description	Status
Antillean Creole	CREOLORAL http://ircom.corpus-ir.fr/site/description_projet.php?projet=CREOLORAL	Audio, Transcriptions, and Translations	Not verifiable
Bastimentos Creole English	Endangered Language Archive http://elar.soas.ac.uk/deposit/0171	Audio, Video, Transcriptions, Translations	Not verifiable; Membership required
Gulf of Guinea Creoles	The Gulf of Guinea Creole Corpora (Hagemeijer et al., 2014)	Document Scans and Transcriptions	Limited Verifiability
Haitian Kreyol	Corpus of Northern Haitian Creole https://www.indiana.edu/~Creole/	Audio and Transcription	Not verifiable

Table 12: Description of Creole datasets presented in our resource survey, which we were not able to verify the existence of. Note here that "Gulf of Guinea Creoles" refers to a collection of four distinct Creole languages: Santome, Angolar, Principense, and Fa d'Ambo. (Part 1/2)

webpage.[6] It is hosted on github pages and allows pull requests so that community members can help us maintain up-to-date information about data available for Creoles.

Moreover, in this section, we would also like to discuss book-based corpora. We do not include them in Tables 12 and 13, as considerable work would need to be done to digitize these datasets, before they can be usable for most NLP tasks. Still, these resources could be useful for those wanting to work with some Creole languages, not listed in Tables 10 and 11 or Tables 12 and 13. Creole corpora documented in books include the Corpus of Written British Creole (Sebba, 1998), a corpus of folktales in Tok Pisin (Slone, 2001), and a corpus of Jamaican Creole (Hinrichs, 2006). We also found the following additional resources described by Kouwenberg and Singler (2009) : transcripts of Guyanese Creole (Rickford, 1987), transcripts of English-based Central American Creoles were introduced by (Holm, 1982), and a corpus of various French-based Creoles, such as Louisiana Creole and Reunionese Creole (Corne, 1999).

NLP FOR CREOLES Creole languages, though largely absent from the NLP literature, have been investigated directly in a small number of works. Of the few works actively focused on Creoles, two works explore directly Creole genesis in the context of computational linguistics. First, Daval-Markussen and Bakker (2012) employ phyloge-

6 https://creole-nlp.github.io/

Language	Resource	Description	Status
Malaccan Portuguese Creole	Endangered Language Archive http://elar.soas.ac.uk/deposit/0123	Audio, Video, Transcriptions, Translations	Not verifiable; Membership required
Mauritian Creole	ALLEX Project http://www.edd.uio.no/allex/corpus/africanlang.html	Concordance of 200k Words	Not verifiable
Nigerian Pidgin	Nigerian Pidgin Tweets (Oyewusi, Adekanmbi, and Akinsande, 2020)	Sentiment Analysis	Not Verifiable
Portuguese Creole	CreolData (Schang et al., 2005)	Lexical Database	Not verifiable
Singlish	Singlish Sentiment Lexicon (Bajpai et al., 2017)	Knowledge Base	Not Verifiable
	Singlish SenticNet (Ho et al., 2018b)	Sentiment Resource	Not Verifiable

Table 13: Description of Creole datasets presented in our resource survey, which we were not able to verify the existence of. (Part 2/2)

netic tools to explore whether Creole langauges form a unique typological group. By treating each Creole as a list of binary linguistic features, including data from WALS (Dryer and Haspelmath, 2013), they analyze the output of a phylogenetic network program (Huson and Bryant, 2006), to inform their investigation. The overall conclusion made by Daval-Markussen and Bakker (2012), was that Creoles indeed formed their own distinct typological class, distinguishable from non-Creoles. However, this work was later refuted by Murawaki (2016), who argued that the study by Daval-Markussen and Bakker (2012) had some methodological shortcomings. Notably, Murawaki (2016) use APiCS features (Michaelis et al., 2013) to encode Creoles, and utilize different approaches for language evolution modeling, to reach the final conclusion that Creoles are *not* typologically distinct from non-Creole languages.

Meanwhile, Lent et al. (2021a) explored the question of how to effectively build language models for three Creole languages (Haitian Kreyol, Singaporean Colloquial English, and Nigerian Pidgin). Their approach involved experimenting with distributionally robust objectives (Oren et al., 2019b), to ascertain whether data from a Creole's "parent" languages could help the language model to be more robust. In the end, they found that straightforward training of language models for Creoles, without adding information from their related languages, produced the strongest results, thus highlighting the relative stability of Creoles.

Finally, there have been a handful of other works aiming to develop NLP algorithms usable for end users, primarily in the area of machine translation, for Creoles like Haitian Kreyol, Mauritian Creole, and Nigeran Pidgin (Ahia and Ogueji, 2020; Callison-Burch et al., 2011; Dabre, Sukhoo, and Bhattacharyya, 2014; Millour and Fort, 2020).

4.3.2 *Notable Features of Creoles*

Many Creole languages are noteworthy for their large capacity for linguistic variation. A speaker's individual style of Creole can vary dramatically depending on social factors, such as their age, ethnicity, geography, and social status. These variations can manifest themselves in different linguistic functions of the Creole, for instance, in the chosen syntax, morphology, or lexical choices (Bajpai et al., 2017). Below, we discuss other features of some (not all) Creoles, that are particularly notable in the context for NLP.

SOCIETAL STIGMA VS RECOGNIZED STATUS Creole languages are infamously stigmatised (Alleyne, 1971; Siegel, 1999). To this day, prejudice against Creole languages has thwarted Creole-based education being made available to Creole speakers, for example. The relative status of a language can change drastically, from Creole to Creole. For instance, use of Singlish has been actively discouraged by government officials, citing the need to "Speak Good English".[7] Meanwhile, a handful of other countries have come to embrace Creole (to varying degrees) in their education system, such as Haitian Kreyol, Papiamento, Seychellois Creole, and Tok Pisin (Kouwenberg and Singler, 2009). The relative celebration or suppression a Creole receives will certainly impact who is speaking the Creole language, and how they will use it.

SPOKEN LANGUAGES Today a large number of Creole languages exist primarily, or almost entirely, as a spoken language only (this can also be a consequence of high stigmatization, as explained in the paragraph above (Sebba, 1997)). If Creole speakers are not typically writing in the language, development of text-based NLP methods may be largely superfluous, unless members of that community have expressed a desire to begin writing (more) in Creole. Consequently, speech technologies may be more relevant to a large number of Creole speaking communities.

NON-STANDARDIZED ORTHOGRAPHY OR GRAMMAR Writing conventions for Creoles can vary greatly, from Creole to Creole, and even from speaker to speaker. Given that Creoles arise from a complex process involving several parent languages (Sessarego, 2020), and for-

7 https://en.wikipedia.org/wiki/Speak_Good_English_Movement

mal writing education in that Creole is not a guarantee for speakers (Siegel, 1999), there is often no standard way of writing them. On one hand, spellings can depend on an individual and informed by their own oral version of the language (Millour and Fort, 2020). Moreover, spelling and grammar conventions in Creole can also be affected greatly by a speaker's proficiency in that Creole. For instance, native speakers of Nigerian Pidgin may speak a fluent, fast, and strong variety of the Creole (i.e., less diluted with English), while others speak a weaker Creole, learned as a second language, characterized by heavy use of just one ancestral Nigerian language. This kind of variety in many cases, as with Nigerian Pidgin, is considered a very positive aspect of a Creole, as it grants speakers a lot of opportunity for nuanced expression. Given that contemporary NLP methods are typically not robust to such linguistic variation, it is important not to limit Creole speakers to one register of communication (Doğruöz et al., 2021).

Meanwhile, some Creole languages are undergoing an ongoing cultural shift, with a push towards standardization, in a manner intended to help cultivate a culture of writing in that Creole. For example, in 2014, a language academy was founded for Haitian Kreyol[8],[9]. For those planning to work on text-based Creole applications, it is vital to become attuned to the current writing culture of that Creole's community, and be aware of how speakers are wanting to use their Creole in writing.

BUGS OR FEATURES? In summary, many of the features discussed above may be perceived as introducing "challenges" or difficult "problems" for NLP to grapple with, as these features are not shared with high-resource languages, like English or Mandarin. However, these so-called "problems for NLP" are often considered positive features by Creole language speakers themselves. We challenge readers not to think of how they can make Creoles work for NLP, but how NLP can work for Creoles.

4.4 WHAT'S WANTED AND WHAT'S NEEDED

In this section, we will give an overview of the key takeaways from our conversations with experts, as well as the major findings from surveying speakers belonging to various Creole speaking communities.

CONNECTING WITH EXPERTS As discussed by Bird (2020), building respectful relationships with the relevant community is absolutely necessary, and reaching out to relevant experts is a great first step towards this direction. For the scope of this work, our definition of

8 https://en.wikipedia.org/wiki/Akademi_Krey%C3%B2l_Ayisyen
9 http://akademikreyol.net/

an "expert" is not strict. We consider an expert to be anyone who is engaged in research, education, or other community outreach, somehow involving the Creole. This can include, for example, individuals working at language schools, field linguists doing research in the area, local scientists in any field, or even graduate students who are native speakers of such languages. Indeed, there are many reasons to begin by reaching out to experts, before even defining your project. First, despite coming from diverse academic backgrounds, experts across different specialities typically speak the same language of scholarship. Although terminological baggage may still interfere with discussion, generally it is easier for fellow field experts to understand, and empathise with each others' goals, than perhaps others. Moreover, even if the experts are not directly working in your field, they may still be familiar or exposed to it. In establishing this relationship, and learning about each other's research or work, there is also a likelihood that some interests overlap, and the opportunity presents itself that you can also help them, which in turn helps to end the norm of treating low-resource language speakers as resources to extract from, and establish a collaborative relationship (Bird, 2020). Additionally, experts also have the authority and knowledge to give you an informed "bird's eye" view of the Creole community, their needs, and desires, as the expert is also a part of it. It's a great (probably necessary) starting point for anyone planning on working on a Creole-language, while not already embedded in the community.

SURVEYING CREOLE SPEAKERS With this in mind, discussion with experts alone runs a large risk of missing out on the thoughts of every day Creole speakers, for whom the language technology is ultimately intended. Thus, their thoughts, opinion, desires, and worries are of utmost importance. For this work, we invited Creole speakers to voice their opinions, and to participate in a survey, through both Twitter and Reddit. Two points should be noted about this approach: (1) One limitation of this method is that our posts already unfortunately exclude Creole speakers not also speaking English or French, and (2) We attempt to break away from extractive/exploitative research practices by asking only those with additional interest in the topic to fill out the survey (i.e., for those individuals without substantial interest, we try to minimize the time required for them to contribute to the discussion, by asking general, open ended questions). While the best case scenario would have been to compensate people for their time, as Hu et al. (2011c) recall, it can be very difficult to find people willing to participate even for payment. Fortunately for this work, we were still able to find a sizeable number of Creole speakers interested in this topic, and willing to have a discussion with us, even if they did not fill out the survey. For the survey, we had 37 participants in total (35 in English, 2 in French), residing in a diverse range of re-

gions (e.g., Caribbean, Africa, North America, Europe, Asia, and the Pacific). We first asked questions about their linguistic background, and use of various languages in daily life. Then, to target NLP wants and needs, we asked more questions about their language use with regards to technology ("e.g. reading/writing SMS on mobile phone, reading/writing on the internet, reading and writing e-mails, interacting with home assistant devices, etc."). For many questions, we included additional prompts, welcoming participants to expand and explain their answers in short-form, which ultimately yielded many important discussion points from the Creole speakers[10].

For the rest of this section, we will review the consistent themes that arose in our conversations with experts, about the wants and needs of Creole language users. These themes will be further expanded upon by the input provided by Creole speakers from our survey. Again, not all themes will be relevant for every Creole. On the contrary, themes seem to be primarily relevant to Creoles with very specific attributes in common (see §4.5).

IS LANGUAGE TECHNOLOGY WANTED OR NEEDED? As discussed throughout this paper, Creole languages are incredibly diverse, including in the way people want (or don't want) to use these languages to interact with technology. Thus, it should come as little surprise, that the answer to the question: "Is language technology wanted and/or needed for this language?", can be everything from "Yes!", "Some technology would be nice", "No", and "Why would you waste time doing that?!", among others.

Amongst both experts and Creole speaking survey respondents, the answer to this question appeared to be largely contingent on how proficient members of the larger community are in the local, high-prestige language (typically English, French, or Portuguese). For instance, a limited subset of the population of Haiti speaks French, and thus Haitian Kreyol is used in most aspects of every day life, and technology to ease the use of Haitian Kreyol is highly desired. On the other end of the spectrum, experts and speakers of Hawaiian Pidgin had difficulties coming up with reasons why language technology support for their Creole would be particularly useful, as the overwhelming majority of speakers (if not all) are highly proficient in English.

CURRENT OBSTACLES In our discussions, some expressed that they already use their Creole for basic tasks, such as texting friends, but that it was not always easy. For example, existing autocomplete or autocorrect software on phones and computers (installed in the relevant high-prestige language, as these technologies are not readily available

10 Please contact us directly if you would like access to our surveys.

to Creoles) often automatically "corrects" Creole spellings or words, and inadvertently suppressing written Creole usage in daily life.

Another issue that Creole speakers mentioned about existing speech technology, was the lack of support for Creole accents or casual code-switching with commonplace Creole words. For instance, navigational assistants for GPS struggle to understand Hawaiian Pidgin accents, in addition to being unable to pronounce local street names, which can be uttered in Hawaiian Pidgin, but not in Standard American English. Extending existing speech technology for dominant, high-prestige languages in this space is much desired, and can be preferable over having a separate, Creole-only system. But without these modifications, existing language technology for the local, high-prestige language can actively harm Creole speakers.

SPEECH TECHNOLOGY As discussed in the Background, many Creole languages are used almost exclusively used in spoken conversation. For such Creoles, text-based language technology are likely moot. Although this can change with time, we encourage readers interested in working in Creole spaces to check with experts and communities, to ascertain if text-based technologies are even needed.

When speech technology was discussed, most Creole languages expressed interest and desire in having speech technology (both text-to-speech and speech-to-text), with the small exception of Creole languages under threat of decreolization (the process by which a Creole ceases to exist), where language revitalization is the dominant concern. But overall, speech technology was perceived by experts and survey respondents to be the most desirable and wanted language technology.

FACILITATING WRITING Some Creole speaking communities already do a lot of writing in their Creole language (despite some obstacles, as we have seen), and/or are trying to foster a culture of writing in the Creole, including standardizing the language. In our conversations with experts and survey responders, we note that there is an expressed need by some Creole communities for basic word processing tools, such as word processors, spell-checkers, grammar-checkers, auto-transcription, etc. However, we found that not all Creole communities welcome *all* of these technologies equally. For example, speakers of Haitian Kreyol mostly welcome spell-checkers, meanwhile speakers of Nigerian Pidgin would eschew these, as it constrains their language use. This point demonstrates how, even when there is a shared desire for a specific kind of language technology, the implementation and specific needs for a Creole can be highly specialized. Lastly, we note that, just as you must learn to walk before you can run, technologies that ease or improve writing in Creoles may

be necessary before Creole speakers could have a need for semantic parsing, for example.

QUESTION ANSWERING AND MACHINE TRANSLATION Both question answering (QA) and machine translation (MT) came up as desired technologies for many Creoles, albeit for different reasons. For Creoles already used online to some extent, QA could improve online search, while MT from Creole into a high resource language, or vice-versa, could provide access to other parts of the world for Creole speakers. Also, MT was cited as desirable for even some endangered Creoles, like Louisiana Creole, as it could help with revitalization. For example, automatic translation from English or French to Louisiana Creole, could allow people to enjoy new domains in Louisiana Creole, and in turn assist with (re)learning the language.

SUMMARY Overall, our discussions with Creole experts and every day Creole speakers underscored how diverse the needs of Creoles can be, for even within one group of languages. We hope this discussion, and the themes put forward, can serve as a springboard for those planning to work on NLP for Creoles.

4.5 CREOLE CONTINUUM FOR LANGUAGE TECHNOLOGY

While the previous section demonstrated that Creoles are not a monolith when it comes to wants and needs for language technology, we did observe several patterns, where Creoles seemed to cluster together, based on their language technology needs, depending on a few shared attributes. To this effect, we introduce a Creole continuum for language Technology (inspired by the post-Creole continuum (DeCamp, 1971)), and propose that there are three key factors that can heavily influence the general needs of a Creole, as follows: (1) Monolingual, Bilingual, or Multilingual community (in other words, is the Creole a lingua franca, facilitating cross-lingual communication?); and (2) General fluency in the relevant prestige language (i.e., do most people also speak the more globally prestigious language, and get on fine, without the Creole?); and (3) Societal acceptance of Creole (e.g., is the Creole language embraced by society as large, or does the Creole struggle from a bad reputation?). We present this continuum in Figure 9, with a small collection of Creole languages, to serve as an example.[11]

The first pattern we would like to draw to the reader's attention to is that the Creoles existing within predominantly monolingual societies, that are also highly fluent in the local prestige language, are

11 We specifically intend the graph axis to be flexible for interpretation, as different Creoles will have different needs, and strict or concrete axis categories may risk reinforcing existing marginalization.

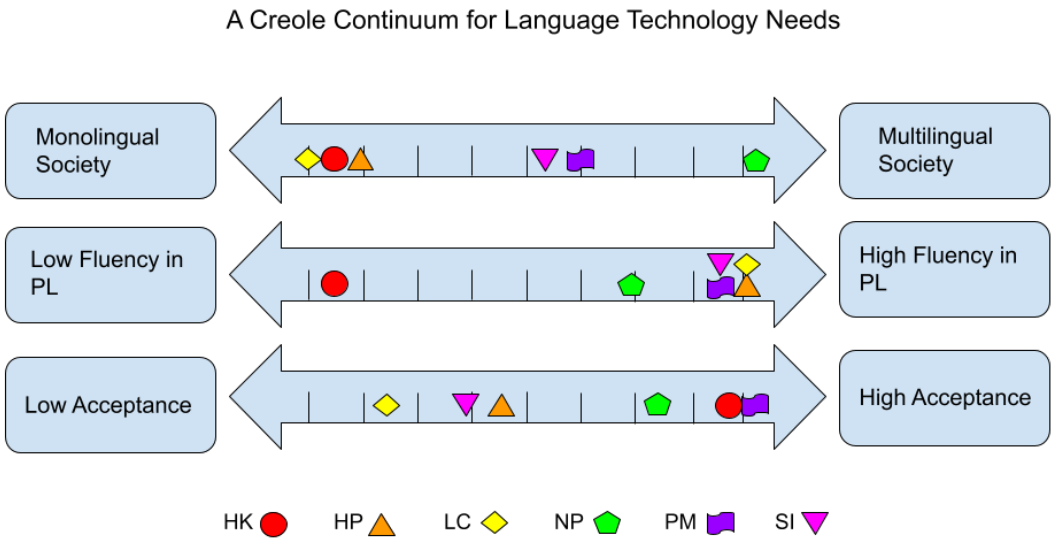


Figure 9: We map a sample of Creole languages to our proposed Creole continuum for language technology. PL here refers to "Prestige Language". We map Haitian Kreyol (HK), Hawaiian Pidgin (HP), Louisiana Creole (LC), Nigerian Pidgin (NP), Papiamento (PM), Singlish (SI).

those that do not need language technology (Hawaiian Pidgin), at least not beyond revitalization (Louisiana Creole). Also, any time that the larger society exhibits very low fluency in the prestige language, language technology is much more likely to be wanted and needed in these communities (Haitian Kreyol). For other Creoles, it is not so clear cut, though. For example, both Singlish and Papiamento exist in generally multilingual societies, with a majority of speakers also fluent in Dutch and English, respectively, and yet the increased societal acceptance of Papiamento (Papiamento is a recognized language of Aruba), means that speakers are more likely to welcome or express needs for language technologies. Still, Singlish is not to be completely neglected, but due to its lower acceptance, language technology suiting more informal situations (e.g. dialog) will likely be more relevant. And finally, the speakers of languages with high acceptance of Creole, namely Papiamento, Haitian Kreyol, and Nigerian Pidgin, are those who typically have the most clear cut wants and needs from language technology, as they already likely use their Creole to interact with technology.

4.6 CONCLUSION

In this work, we have demonstrated that Creole languages should be of larger interest to the NLP community, and we provide a survey of resources and NLP research produced for Creoles. In doing this, we

have also shown that Creoles cannot be conflated together, if we are to make language technology that is truly useful for a community. Truly, the best approach to developing NLP for Creoles to to get in contact with both experts and community members, and listen earnestly to their wants and needs for language technologies, as well as what is specifically not wanted.

5

TOWARDS CREOLE NLP

5.1 ABSTRACT

Benchmark datasets serve as a common resource for researchers developing new and improved language technologies for a given task or language. Such datasets enable the field to easily track progress in an area over time, and can also serve as a convenient entry point for researchers who have not previously worked in that space. In this chapter, we describe our ongoing work towards creating a multitask, multilingual benchmark dataset for Creoles, a category of low-resource languages that are still largely absent from the current NLP landscape. In creating and assembling such a resource, we hope to encourage NLP researchers to include Creoles in their work, and provide data that can be helpful in developing language technologies intended for Creole-speaking communities.

5.2 INTRODUCTION

Benchmark datasets have played a key role in advancing performance of many tasks across NLP such as question-answering (Bartolo et al., 2020; Rajpurkar et al., 2016a; Yang et al., 2018), semantic parsing (Hershcovich et al., 2019; Yu et al., 2018; Zhong, Xiong, and Socher, 2017), and dialog (Budzianowski et al., 2018; Eric and Manning, 2017), as well as encouraged progress for low-resource languages (Doan et al., 2021; Guzmán et al., 2019) and specialized domains like law and finance (Chalkidis et al., 2022; Chen et al., 2021). Multitask benchmarks have also allowed for assessment of new approaches over a wide variety of tasks at once (Wang et al., 2019, 2018), and at the same time multilingual benchmarks have enabled comparison of performance over a collection of languages (Adelani et al., 2021; Agić and Vulić, 2019; Conneau et al., 2018; Hu et al., 2020; Nivre et al., 2020b). Multilingual benchmark datasets in particular have also been vital for observing successes and failures of approaches to cross-lingual transfer (Artetxe, Ruder, and Yogatama, 2020; Lewis et al., 2020b; Vries, Wieling, and Nissim, 2022).

Unforutunately, most Creole languages are still "left behind" when it comes to benchmark datasets (Joshi et al., 2020b). Presently very few datasets exist for individual Creole languages, and existing datasets typically pertain to one individual task (Lent et al., 2022), and Creoles as a whole are are poorly represented within existing multilingual benchmark datasets, if represented at all. The consequences of this

vacancy are two fold: 1) development of language technology continues to lag behind for Creoles, and 2) conclusions about cross-lingual learning are being drawn without representation of Creoles. The latter is concerning, as recent evidence suggests that common assumptions about transfer learning (i.e., that it should be typically possible to achieve reasonable transfer from related languages) might not be trivially applied to Creoles (Lent, Bugliarello, and Søgaard, 2022). Although Creoles are not the focus of their work, Vries, Wieling, and Nissim (2022) provide further evidence that ancestor-to-Creole transfer may be difficult, as poor accuracy (below 50%) on UDPoS was observed for Nigerian Pidgin as the target language, after fine-tuning XLM-RoBERTa base (Conneau et al., 2020) on either English or Portuguese as the source language.

5.2.1 *Contributions*

In order to assist development of language technology for Creole languages, as well as facilitate further investigations of cross lingual learning, we are working towards `CreoleGLUE`, a multitask benchmark dataset for a collection of Creole languages, which will be released to the public upon completion. This new benchmark dataset will encompass both new (§5.3, §5.4, §5.5) and existing (§5.3.2 §5.6) datasets, for a wide variety of Creoles and NLP tasks. In this chapter, we also contribute a discussion of challenges and considerations for low-resource data collection, including cultural relevance of data (§5.3.1 §5.4), as well as a description of our plans for finalization of the dataset and running benchmark experiments (§5.3.2, §5.4, §5.5, §5.6).

5.3 CREOLE WIKI

Both Wikipedia and Wikidata have long been key resources leveraged by the NLP community. Wikipedia has been used to train large language models (Devlin et al., 2019), and also for generating datasets for a multitude of tasks, such as name entity tagging (Althobaiti, Kruschwitz, and Poesio, 2014; Littell et al., 2016; Nothman, Curran, and Murphy, 2008), entity linking (Lin et al., 2016; Pan et al., 2017), text summarization (Fatima and Strube, 2021; Zopf, Peyrard, and Eckle-Kohler, 2016), and question answering (Kwiatkowski et al., 2019; Liu et al., 2020). Meanwhile, Wikidata is often used in database question answering (Cao et al., 2022; Cui et al., 2021; Diefenbach et al., 2017; Korablinov and Braslavski, 2020; Saha et al., 2018), but also has demonstrated utility within other tasks such as named entity recognition (Nie et al., 2021), entity linking (Kannan Ravi et al., 2021), and coreference resolution (Aralikatte et al., 2019).

Creole	Wiki Code	Num Pages	Num Usable Pages	Avg Toks	Med Toks	Max Toks
Haitian Creole	ht	70778	61728	34	16	15193
Chavacano	cbk-zam	4719	3300	94	27	15242
Guyanese Creole	gcr	2397	2383	75	38	2506
Papiamento	pap	2903	2375	148	34	40561
Jamaican Creole	jam	2281	1741	86	42	7307
Tok Pisin	tpi	1977	1583	26	10	1305
Bislama	bi	1698	1482	22	15	1107
Piktern	pih	1229	922	28	17	700
Sango	sg	666	331	24	14	1641

Table 14: Statistics on the 9 Creoles with available Wikipedia dumps. Wikipeda's language codes are listed here, as they do not necessarily match ISO-3 codes. Num Pages indicates the base number of Wikipedia Pages included in a dump, but Num Usable Pages indicates the number of unique (i.e. excluding duplicates), non-empty pages. Avg Toks, Med Toks, and Max Toks indicate the average, median, and maxiumum number of tokens (split on white space) within a Creole's Wikipedia dump. All non-empty pages had at minimum 2 tokens across all Creoles.

Presently, Wikipedias exist for over 327 languages[1], of which we identified 16 Creole languages in total. To start, we downloaded all available dumps[2] of the Creole Wikipedias, as listed in Table 14. Unfortunately, only 9 of the 16 Creoles had available dumps, leaving 7 Creoles which still remain to be scraped[3]. The Wikipedia dumps were processed and cleaned using WikiExtractor[4], and then further filtered by removing duplicate and empty pages. In Table 14, we observe that this simple filtering heuristic can at times greatly reduce the number of usable pages (i.e., the difference between "Num Pages" and "Num Usable Pages"), demonstrating that naive page count alone cannot give an accurate account of how much data there is to work with from a given Wikipedia. Furthermore, if we look into the median number of tokens ("Med Toks") for each article in a Creole Wikipedia, we can observe that, for some Creoles, a typical Wikipedia page likely contains just one or two very short sentences (e.g. Tok Pisin or Sango).

After filtering the dumps, we used the Wikidata API to link each article's main entity (i.e. the title) to its associated entity code (i.e. Qcode). Because Wikidata exists as large knowledge graph with detailed taxonomies about entities, linking texts to their primary Qcode enables us to repurpose the data for several NLP tasks for the Creoles with Wikipedias. For example, as we can see in Figure 10, having the Qcode for an article allows us to access subsequent information about

1 https://en.wikipedia.org/wiki/Wikipedia
2 https://dumps.wikimedia.org/backup-index.html
3 Cape Verdean Creole, Guadeloupe-Martinique French, Krio, Mauriitian Ceole, Pijin, Reunion Creole, and Seychellois Creole
4 https://github.com/attardi/wikiextractor

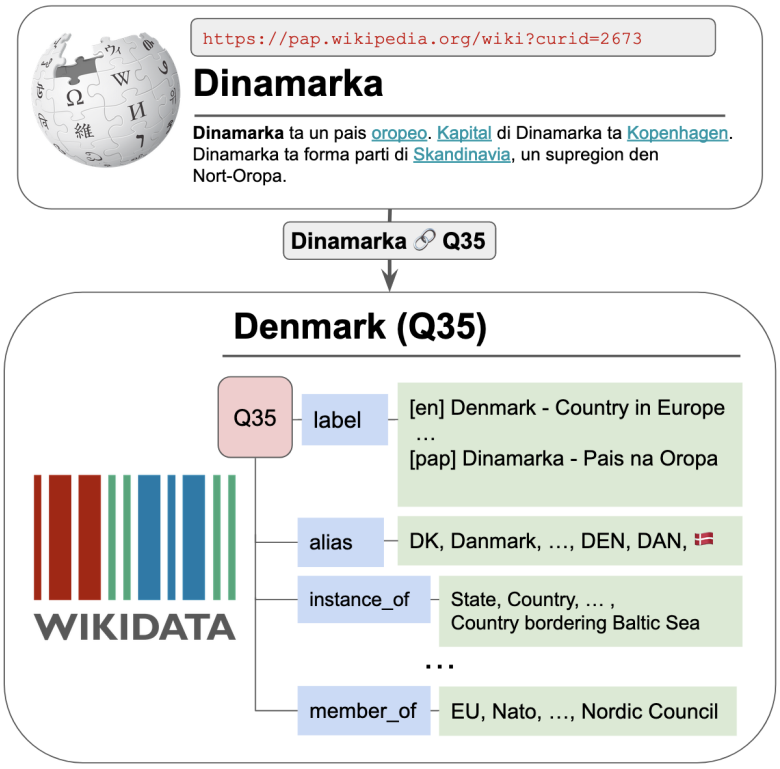


Figure 10: Example of a linking an article's main entity from the Papiamento Wikipedia to a Wikidata.

that entity's label (including descriptions in many languages), other names that also refer to the same entity (i.e. alias), and we can find many relational properties of an entity, such as if it becomes to any organizations (i.e. member_of). Depending on what kind of entity it is (e.g. is it an instance_of a State versus a musical group?), different information about properties will be available. For `CreoleGLUE`, we identify two tasks, for which we can readily create evaluation benchmarks for: **document classification** and **paraphrasing**. For document classification, we can take entities from the "instance_of" relationship, and use these as class labels for the source document. For instance, the text in the Wikipedia page about Denmark in Figure 10, could be given a label "Country", as could a text about Elton John be given a label "Person". Meanwhile, for paraphrasing, we can utilize information about the main entity's label and alias, to create paraphrases. Though, this requires that there must aliases and labels available in the pertinent Creole, which has not yet been verified. A more thorough paraphrasing dataset could be made if more entities in the Wikipedia text were also linked to Wikidata, than just the main entity. This process of complete linking to Wikidata is still in progress (see §5.3.2).

5.3.1 *Pitfalls of Wikipedia for Lower-Resourced Languages*

While Wikipedia is a commonly used resource across NLP, Wikipedias for different languages can vary dramatically in terms of size (i.e. number of articles), but also *quality*. And while NLP research conducted on scores of languages are important for promoting linguistic diversity (), we believe that quality of Wikipedia data for these large-scale studies is too often overlooked. In this work, we performed a cursory quality check of the Creole Wikipedias, and identified three reoccurring pitfalls within the data: (1) frequent use of templates, (2) multilingual noise, and (3) outdated or less relevant content. We explain these pitfalls in more detail below, and as a result of these shortcomings in the data, draw the conclusion that researchers must not take for granted the disparities in data quality between higher- and lower-resourced data, in the context of large multilingual NLP studies. To better understand these quality discrepancies, we believe an interesting and important area for future work would be the development of methods to automatically evaluate the quality of Wikipedias for different languages.

TEMPLATES During our qualitative assessment of the Creole Wikipedias, we found many pages to be composed of examples drawing from apparent templates. The which raises concerns whether the sentences were generated, and in turn, to what extent they sound natural. Below are several examples from the Bislama Wikipedia pertaining to geography, which demonstrate the ubiquity of templates:

(1) Arizona i wan state blog Yunaeted Stet. Kapital blong hem i Phoenix. Long July 2009, populaesen blong Arizona i stap araon 6,931,071.

(2) Alaska i wan state blog Yunaeted Stet. Kapital blong hem i Juneau. Long July 2016, populaesen blong Alaska i stap araon 741,894.

(3) Honolulu hem i kapital blong Hawaii. Long July 2016, populaesen blong Honolulu i stap araon 351,792.

(4) Rome hem i kapital blong Itali. Hem i stap long Lazio rijon.

Even without understanding Bislama, a human can look at the above examples and easily identify the following templates:

- `STATE` i wan state blong `COUNTRY`. (Matches # 1, 2)
- Kapital blong hem i `CITY`. (Matches # 1, 2)
- `CITY` hem i kapital blong `PLACE`. (Matches # 3, 4)

- Long DATE, populaesen blong PLACE i stap araon POPULATION. (Matches # 1, 2, 3)

While template-generated language has been demonstrated to be useful as auxiliary data on a number of tasks (Athreya et al., 2021; Lent et al., 2021b; Wang, Berant, and Liang, 2015; Yu et al., 2021), having templatic data as a core part of your primary training and/or evaluation data is worrisome for several reasons. For example, a language model trained on such examples will be brittle to more natural-sounding paraphrases, especially to utterances less resemble the templates.

MULTILINGUAL NOISE Wikipedias for lower-resourced languages can be surprisingly noisy with multilingual data. For example, a sizeable category within the Haitian Creole Wikipedia is composed of pages for Spanish-speaking actors, and these actors' pages often only contain a list of their Spanish-titled films[5]). As a result, some part of the Haitian Wikipedia is entirely Spanish. While this non-Creole noise can be fairly straight forward to clean out of the dumps when scripts are different (e.g. identifying and removing Chinese scripts from the Haitian Wikipedia), removing this noise becomes increasingly difficult for more closely related languages (e.g. French words in Haitian Wikipedia, or English words in Jamaican Creole Wikipedia), where a word might be shared by both languages. Moreover, most Creole languages are still not represented in popular, publicly available language identification models. Thus, the "Num Usable Pages" column in Table 14 is likely still an inaccurate depiction of how much data there truly is to work with for a Creole, as pages consisting largely of non-Creole still need to be cleaned out, and this is not trivial.

NOTES ON CONTENT Finally, in our qualitative evaluation of Creole Wikipedia quality, we found a few notable issues about content within the Wikipedias. We note, first, that some inaccurate content can be found within the Creole Wikipedias, and second, that some of the Creole Wikipedias contain larger amounts of religious content than others.

To start, for an example of inaccurate content, the Haitian Wikipedia page about Haitian Creole[6] contains outdated information about the Haitian alphabet[7]. We hypothesize that there may be a correlation between Wikipedia size and content accuracy – with a smaller Wikipedia, fewer people are able to use it as a practical resource for finding information, which also means there will be fewer people maintaining and

5 https://ht.wikipedia.org/wiki?curid=65052
6 https://ht.wikipedia.org/wiki/Krey%C3%B2l_ayisyen
7 The page incorrectly discusses "lòt òtograf", which is not a part of today's standardised Haitian alphabet, which can be found here: https://mit-ayiti.net/resous/an-n-konprann-chante-alfabe-kreyol-la/

updating the pages. This is in contrast to larger Wikipedias, like that for English, where many people can use it as a practical resource, and thus more people find and correct pages with outdated information, resulting in a more up-to-date resource. Of course, if a Wikipedia is largely templatic (see §5.3.1), these kinds of content inaccuracies may be less present, even when a Wikipedia is small.

Another potential issue with content in the Wikipedias is an abundance of religious content for the Sango and Tok Pisin Wikipedias. For example, the Tok Pisin Wikipedia page about biological natural selection[8], discusses the biblical story of Adam and Eve. We hypothesize that over representation of religious content, in comparison to the other Creole Wikipedias we looked at, may be more likely to occur in Wikipedias, for languages with speakers who by and large do not have access to the internet (e.g. Sango in Central African Republic and Tok Pisin in Papua New Guinea[9]). In these communities, missionaries of various faith groups may be more likely to participate in the push towards digitization, resulting in content that more closely resembles or echos other religious texts. However the specific verbiage used in the bible or other religious texts is known to be less ideal for NLP applications intended for end users, because it is typically old-fashioned or otherwise misaligned with the actual linguistic habits of speakers (Agić et al., 2016; Mielke et al., 2019; Östling and Tiedemann, 2017) and the narrow domain .

Ultimately, without having native speakers evaluate each Creole Wikipedia, it is impossible to say how pervasive these content concerns are, and to what extent they pose a problem for Creole NLP. Still, these issues are important for NLP practitioners to be aware of, as we know that poor quality and irrelevant content in the training data can contribute negatively to considerable bias(Bender et al., 2021; Blodgett et al., 2020; Bolukbasi et al., 2016).

5.3.2 *Next Steps*

Before any data from Creole Wikipedia is ready to be included in a benchmark dataset intended for the public, there are a number of necessary steps we must act on. First, we will integrate Wikidata with the full Wikipedia articles, and not just the title entities. This will allow us to perform a feasibility check for the paraphrasing task, as we will verify that there are enough labels and aliases for the Creoles in Wikidata. Once this is done we can subsequently generate the paraphrased data.

Next, for both document classification and paraphrasing, we must create the evaluation files. Due to the extremely small sizes of the

8 https://tpi.wikipedia.org/wiki?curid=8835,Netirel%20seleksan

9 https://en.wikipedia.org/wiki/List_of_countries_by_number_of_Internet_users

Creole Wikipedias (Table 14), it may not be advisable to also create training data for every single Creole. The most realistic setting for work on Creoles is in the context of zero-shot learning, as Creole data is extremely low-resourced, and Creoles are also not represented in publicly available, large-scale multilingual pre-trained langauge models, such as mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020), although the pre-training data for mT5 (Xue et al., 2021) is estimated to consist of 0.33% Haitian Creole data (for context, the training data is estimated to consist of 107 languages, and 5.67% of this data is in English).

Finally, we would like to integrate WikiAnn (Pan et al., 2017), a benchmark dataset for named entity recognition and entity linking within Wikipedia articles, into `CreoleGLUE`. Of the Creoles listed in Table 14, only Haitian Creole, Papiamento, Tok Pisin, Bislama, Pidgin, and Sango are present in WikiAnn, and all other Creoles absent. As part of this integration, we will perform a quality assessment of the examples; because WikiAnn is "silver-standard" generated data, we believe that the Creole subset of WikiAnn may also be negatively affected by some of the pitfalls, as described in §5.3.1.

5.4 MACHINE COMPREHENSION FOR CREOLES

Machine comprehension (MC) is a subtask within question answering (QA), which aims to produce a model that can reason over text. In order to correctly answer questions about a given text, the model should understand the setting, actors, and events, and how they relate to one another. Presently, datasets for MC exist for predominantly high-resource langanguage, such as English (Rajpurkar et al., 2016b; Richardson, Burges, and Renshaw, 2013; Trischler et al., 2017; Welbl, Stenetorp, and Riedel, 2018) and Mandarin (Yang et al., 2018), due to the difficulties and large expenses in creating these datasets. In this work, we hire translators to create the first over Creole MC evaluation datasets, by having a hallmark MC dataset, MCTest (Richardson, Burges, and Renshaw, 2013), translated into Marutitian Creole and Haitian Creole. Additionally, as it is well known that translationese can reduce a models' cultural relevancy for its target audience, Hershcovich et al., 2022; Lent et al., 2022, we present the first ever *cross-cultural* MC dataset, including two separate translations into Haitian Creole – one standard and one localized dataset. We believe this will be the **first published dataset enabling explicit development of cross-cultural NLP**.

TRANSLATING MC TEST MCTest is a publicly available dataset, composed of reading comprehension questions over stories appropriate for similarly testing the reading comprehension skills of young, school-aged children (i.e. 7 years old) (Richardson, Burges, and Ren-

shaw, 2013). The training dataset is in English, and consists of either 500 stories (MC500) or 160 stories(MC160), and in both setups, each paired with 4 multiple-choice questions, which require reasoning over the story text to correctly answer. Within the dataset, questions are labeled for whether they require reasoning over multiple sentences versus one sentence, to answer correctly. Due to the high financial cost of hiring translators, we chose to translate the development data for MC160, which consists of 30 stories, and in total contains 120 questions. As Hu et al. (2011c) note in their work, it can be very difficult to find workers in lower-resourced languages, despite offering payment. In the end, we were able to hire 2 professional translators, one speaking Mauritian Creole, and the other Haitian Creole, to obtain our three translationsThe cost of each individual translation was roughly a thousand euro, to give readers an idea about the cost of such work.. With stories and questions combined, a single MC160 development contains approximately 10,000 words. While a full translation of the full MC160 dataset would have been ideal, believe translation of the development set is a good start, and will be sufficient for initial work on Creole machine comprehension.

LOCALIZATION For both Mauritian Creole and Haitian Creole, we obtained standard translations of the MC160 development set. Here, "standard" means that the translators were instructed to keep translations as true to the original documents, without compromising grammar and fluency, when sentences need some re-wording to sound natural. Fundamentally, the meaning and content are the same, though. Standard translations, without changes to the core content, are necessary to directly compare performance with the source data, and measure transferability concretely.

Additionally, we were also able to receive a localized translation for Haitian Creole. In contrast with a standard translation, localized translations will modify content that is not relevant for the pertinent culture, into something else that is both relevant for that culture, but also still matches the context and scope of the text being translated. A simple demonstration can be found in Figure 11. To start, the text was first translated into Haitian Creole with the standard approach. We can see that the highlighted names ("Greta" and "Tony") and entities ("ice cream truck" to "*kamyon krèm*") are preseved in the standard translation for Haitian, despite the names not being Haitian, and a ice creams trucks not existing in Haiti. For the localized translation, the names and entities need to be relevant for Haitians. Thus the name "Greta" is changed to a Haitian name, "Agat", and "Tony" has been changed to "Toni", to make the name sound more Haitian (i.e. with Creole spelling). Likewise, the ice cream truck ("*kamyon krèm*") is removed, and replaced with a similiar, Haitian entity: the "*machann fresko*", which is a traveling vendor with a cart, who sells various

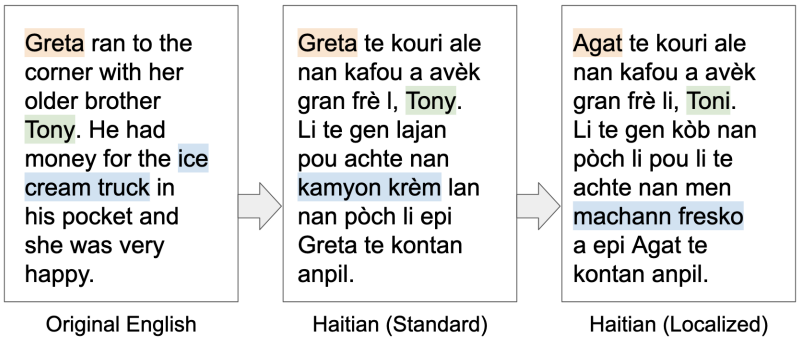


Figure 11: Example of standard and localized translations of into Haitian Creole from English, with highlighting of notable entities, which are changed between the standard and localized translations.

kinds of treats. As such, any mentions of "ice cream" in the story are changed into a "*fresko*", a shaved-ice treat enjoyed with different syrups.

Other kinds of changes made for localization can include (but are not limited to) **activity changes** (e.g., baseball is not played in Haiti, so it can be a changed into a marble game, *jwèt mab*), **location changes** (e.g., a story set near the "Chato Hood River" has been localized to "Chato Gran Rivyè", as this keeps closer to a familiar or historical place in Haiti), and **animal changes** (e.g. famous animals like lions, tigers, and bears are kept, as Haitians of course also know them, but lesser known animals are changed to local ones).

In the few case where a thorough localization would require rewriting the entire story (e.g. something very involved with the activities, like in MCD160 development set story #8, which asks detailed questions about playing baseball), the non-local activity is maintained, and localization is done through other techniques, such as those that we have discussed already. It is not unforeseeable that school aged children (the target human audience of MCTest) could learn about sports in other countries in school, and thus we do not perceive these few instances of more conservative localization to be harmful, as it also still allows us to compare against the standard translations and original English data.

ERRORS IN MCTEST During the translation process, our translators found and corrected a two small errors in MCTest. We will release a corrected English version of the MC160 development set, to match our correct Creole translations, with clear documentation about our updates, when the data is released to the public. The first error was for `mcd160.dev.3`, where question #3 should remove "and Greta" from the correct answer; the second error was for `mcd160.dev.17`, in which the story discusses a yellow flower, and the correct answer incorrectly uses "pink". These questions have been fixed across a copy of the En-

glish data, the Mauritian Creole translation, and both Haitian Creole translations.

NEXT STEPS Our remaining tasks for this work are to run benchmark experiments for the English data, as well as our 3 Creole datasets. We should compare the zero-shot learnability when using pre-trained language models that have *not* seen Mauritian or Haitian Creole (i.e. mBERT and XLMR Conneau et al., 2020; Devlin et al., 2019), and a pre-trained language model that has seen Haitian Creole (i.e. mT5 Xue et al., 2021). Moreover, a very important comparison will be that between Haitian-Standard and Haitian-Localized. To what extent (if any), there is a delta in performance between the two, will be very interesting, and a first step towards evaluating cross-cultural NLP in a concrete, quantitative setting. Before the data is released, we will also include a thorough documentation, with information about localization and corrections made, for the translations.

5.5 MIT HAITI CORPUS

As explored in §5.3.1, Wikipedia data for low-resourced languages is not always guaranteed to be the highest quality. In efforts to assemble a very high quality dataset for Haitian Creole, we have begun efforts to build a new corpus of posts, lesson plans, and other documents from MIT Haiti[10], an initiative that helps to design and desiminate educational materials for Haitian students in Haitian Creole, with special attention paid to the STEM fields and encouraging active learning. With the permission of the organization, we scraped the website `https://mit-ayiti.net/`, which consists of high quality educational materials, both in plain text and PDF's, paired with a meticulous tagging system, so that each blog post or lesson plan has ample metadata about the subject. For example, a page will be tagged with the subject(s) (e.g. "*Jeyografi*" for Geography), and also the intended grade-level for the material (e.g. "*Preskolè*" or "*3e Ane Fondamantal*"). While we are still in the early stages of assembling this corpus, this extensive tagging system will allow us to, at minimum, create a high quality evaluation dataset for document classification, if we will map the tags to a similar set of classification labels, within the scope of a pre-existing task. For plain text documents scraped from the site, thusfar we have 190 unique pages with extensive tags. As the bulk of the MIT Haiti materials exist as PDFs, our next step is to download the PDFs, and try to use a PDF-to-plain-text converter, to see if we can successfully obtain the PDF text, too. This would substantially increase the amount of data we have to work with, and thus expand the realm of possibilities of how we can work with this data.

10 `https://haiti.mit.edu/`

NEXT STEPS Beyond trying to leverage the high quality data within the PDFs, there are a number of other steps we must take, before we can prepare the data for document classification, or any other tasks. The first, is to investigate the distribution of tags over the documents that we have thusfar, to get an overall picture of which categories are best represented, which categories most often overlap, etc. Because the tagging system within MIT Ayiti is so thorough, it will hopefully be possible to create multiple document classification tasks - one based on the subject (e.g., physics, mathematics, literature), and another based on the grade level (e.g., preschool, 1st grade, and so on). For the case of the latter, such a categorization of this data could be interesting for studying applications of text simplification (Van, Tang, and Surdeanu, 2021), or for studying techniques in curriculum learning (Cirik, Hovy, and Morency, 2016), where easier examples (i.e., examples intended for younger students) are shown to the model before harder ones. Indeed, there is great potential for this dataset, as the source material is very high quality, and much work remains to be done on curating this resource.

5.6 OTHER BENCHMARKS

Thus far in this work, we have mostly described our own efforts in collecting and translating data for Creole NLP, with the exception of WikiAnn Pan et al., 2017 discussed in §5.3.2. Yet there are also several other published datasets for individual Creoles, which we would like to include in `CreoleGLUE`, both to expand on the number of Creoles, as well as the number of tasks.

TASKS FOR NIGERIAN PIDGIN Of all of the Creole languages, Nigerian Pidgin English (often simply called "Nigerian Pidgin" or even just "Naija") has received the most attention in recent years, with a number of new dataset releases Lent et al., 2022, and we would be amiss to not include these into `CreoleGLUE`. The first notable dataset is `MasakhaNER` Adelani et al., 2021, which is a multilingual named entity recognition benchmark for African languages, including Nigerian Pidgin. Results on this dataset will not be directly comparable to the named entity recognition for the Creoles in WikiAnn Pan et al., 2017, as those are more akin to "silver" data, but can rather show us, perhaps more accurately, how Creole NER is performs on "gold" data (Nigerian Pidgin is also absent from multilingual pre-trained language models, so the methods will be comparable across Creoles).

Two other tasks for Nigerian Pidgin datasets that we'd like to include in our benchmark are `NaijaSenti` (Muhammad et al., 2022), a sentiment analyis task over tweets, and `SUD Treebank for Naija` (Caron et al., 2019), a Universal Dependency parsing task. While there are presently no opprotunities to add sentiment analysis or depen-

Creole	Document Classification	Paraphrasing	NER	MC	UD	Sentiment Analysis
Bislama	✓	✓	✓			
Chavacano	✓	✓				
Guyanese Creole	✓	✓				
Haitian Creole	✓✓	✓	✓	✓✓		
Jamaican Creole	✓	✓				
Mauritian Creole				✓		
Papiamento	✓	✓	✓			
Nigerian Pidgin			✓		✓	✓
Piktern	✓	✓	✓			
Sango	✓	✓	✓			
Singlish					✓	
Tok Pisin	✓		✓			

Table 15: Anticipated NLP tasks and Creole languages to be include in the benchmark dataset. Here, a checkmark indicates that a dataset should be available for evaluating performance on the Creole for the specified task.

dency parsing for the other Creoles discussed thusfar in this work, these addition of these datasets will allow us to expand the breadth of tasks within `CreoleGLUE`.

A TASK FOR SINGLISH Singlish, formally known as Singaporean Colloquial English, is a mostly spoken Creole from Singapore (Lent et al., 2021a). As such, there are limited text resources available. However, we would like to include the `Singlish Dependency Treebank` (Wang, Yang, and Zhang, 2019), for the addition of another Universal Dependency parsing task over a Creole (in conjunction with Nigerian Pidgin).

SUMMARY By including these other benchmark datasets into `CreoleGLUE`, we aim to have a wide range of NLP tasks over a diverse group of Creole languages. In Table 15, we present a summary of our vision for `CreoleGLUE`. For Haitian Creole, there are two checkmarks for both document classification and machine comprehension, as we expect to have document classification datasets derived both from Wikipedia and MIT Haiti, and as there are two translations (standard and localized) for MCTest.

5.7 CONCLUSION

This chapter documents the progress made so far towards `CreoleGLUE`, a multilingual, multitask benchmark dataset with new and existing

tasks for Creole NLP. New tasks include document classification, paraphrasing, and cross-cultural machine comprehension, while datasets in named entity recognition, entity linking, and sentiment analysis already exist in some Creoles, and will be incorporated into our benchmark dataset. In describing each of these tasks, we have also outlined our plans and vision for this project (e.g. "Next Steps"), and also provided a thorough discussion of various topics important to dataset creation (e.g., pitfalls of Wikipedia and translation localization). We hope that this chapter will excite readers about this benchmark dataset, which we believe will be important for improving studies on multilingual and cross-lingual NLP, as well as bringing the field one step closer towards Creole NLP.

Part III

EVALUATION OF SEMANTIC PARSERS

6

TESTING CROSS-DATABASE SEMANTIC PARSERS USING CANONICAL UTTERANCES

6.1 ABSTRACT

The benchmark performance of cross-database semantic parsing has climbed steadily in recent years, catalyzed by the wide adoption of pre-trained language models. Yet existing work have shown that state-of-the-art cross-database semantic parsers struggle to generalize to novel user utterances, databases and query structures. To obtain transparent details on the strengths and limitation of these models, we propose a diagnostic testing approach based on controlled synthesis of canonical natural language and SQL pairs. Inspired by the CheckList (Ribeiro et al., 2020), we characterize a set of essential capabilities for cross-database semantic parsing models, and detailed the method for synthesizing the corresponding test data. We evaluated a variety of high performing models using the proposed approach, and identified several non-obvious weaknesses across models (e.g. unable to correctly select many columns). Our dataset and code are released as a test suite at `github.com/hclent/BehaviorCheckingSemPar`.

6.2 INTRODUCTION

Cross-database semantic parsing, the task of mapping natural language utterances to SQL queries for any database, has attracted increasing attention since the introduction of benchmarks like WikiSQL (Zhong, Xiong, and Socher, 2017) and Spider (Yu et al., 2018). The advent of pre-trained language models (Devlin et al., 2019; Lewis et al., 2020a; Liu et al., 2019; Peters et al., 2018) has further accelerated the progress in this area (Choi et al., 2020; Lin, Socher, and Xiong, 2020; Shi et al., 2020; Wang et al., 2020; Yu et al., 2020).

Despite impressive gains on standard benchmarks, studies on cross-database semantic parsing models show that they still suffer from out-of-distribution (OOD) generalization when presented with novel user utterances (Radhakrishnan, Srikantan, and Lin, 2020; Shaw et al., 2021; Suhr et al., 2020), databases (Suhr et al., 2020) and SQL query structures (Finegan-Dollak et al., 2018; Shaw et al., 2021; Suhr et al., 2020). As baseline performance climbs ever upward, at what point can we confidently deploy our models to end users, and how will we know we have reached this point?

Inspired by Ribeiro et al. (2020), which has shown the effectiveness of simple, systematic, and heuristic behavior checking strategies

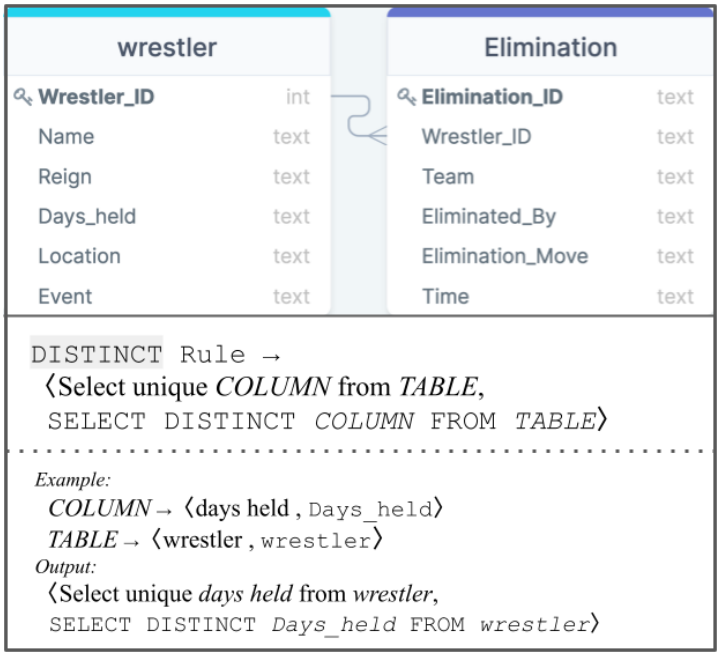


Figure 12: The database (top) is applied to our SCFG production rule (middle) to produce a new example for the `DISTINCT` category (bottom). See Appendix A.5 for production rules of other categories.

for evaluating the robustness of NLP models, we propose a controllable, non-adversarial unit testing approach to shed more light on the capabilities of cross-database semantic parsers. We implement a synchronous context-free grammar (SCFG) to generate natural language questions based on SQL queries (Figure 12). This grammar features production rules that evaluate important categories of SQL element types such as clauses (e.g. `SELECT` and `WHERE`), as well as commonly used operators including aggregators (`MAX`), conditionals (`BETWEEN`), and logical operators (`OR`). We handcraft the rules for these categories to ensure that the generated question-query pairs are simple, natural, unambiguous, and with minimal cross-category overlap.

We apply our evaluation framework to four state-of-the-art text-to-SQL models, namely BRIDGE (Lin, Socher, and Xiong, 2020), RATSQL-RoBERTa and RATSQL-GraPPa (Yu et al., 2020), and RATSQL-GAP (Shi et al., 2020), and observe that these models struggle to extend their success on the Spider dev set consistently to our evaluation data, with the exception of a few categories. Further analysis of the fine grained categories shows that they also fail on many rudimentary test cases (e.g., selecting multiple columns and properly producing conjunctions). While existing studies show that the models tend to fail on challenging cases that involve novel user expression (Suhr et al., 2020) and SQL structures (Shaw et al., 2021; Suhr et al., 2020), our diagnosis exposes more robustness issues in their surface form understanding (even with seemingly simple inputs), and highlights the importance of addressing such issues in the modeling foundation (Bommasani

et al., 2021). Our dataset and code are released as an extensible test suite.

6.3 RELATED WORK

PARAPHRASING A number of augmentation methods have been made to create paraphrases of the input query, with methods such as synonym replacement (Kwiatkowski et al., 2013), use of a paraphrase model (Berant and Liang, 2014), and backwards utterance generation (Zhong et al., 2020). While these approaches ensure the creation of additional examples with more variation on the natural language side, they can be vulnerable to error, when a wrong synonym or paraphrase is chosen by a model. Although such errors may amount to just noise when used as additional training data in conjunction with a benchmark dataset, they make evaluation on such generated sets impossible, unless examples with errors are manually removed from the dataset.

CANONICAL UTTERANCES Wang, Berant, and Liang (2015) demonstrated that it is possible to lessen the reliance on humans for creating a dataset by first generating logical forms and canonical utterances, and then use crowdsourcing to create more natural-sounding paraphrases of the questions. They note that this method is particularly effective when you seek to quickly create data for creating a *domain specific* parser. Iyer et al. (2017) also demonstrated that crowdsourced annotations from such approaches, as in turn user feedback in an online setting, can be used improve parses and detect incorrect queries. Although originally designed in the context of transfer-based machine translation to generate translation pairs (Chiang, 2005), SCFG's have also been adapted in previous semantic parsing work (Wong and Mooney, 2006, 2007) for generating new sentence-parse pairs. More recent utilization's of SCFG's for semantic parsing induce the grammar and use the resulting data for additional training and pretraining (Jia and Liang, 2016; Yu et al., 2020).

ROBUSTNESS TESTING Finally, Ribeiro et al. (2020) has demonstrated the efficacy of handcrafting templates for generating data points to "unit test" the models. We design synchronous context-free grammar (SCFG) production rules to generate test data for specific cross-database semantic parsing capabilities. Other NLP evaluation frameworks that look beyond accuracy and target a more general set of NLP tasks have also been proposed (Goel et al., 2021; Kiela et al., 2021; Liu et al., 2021b).

6.4 GENERATING CANONICAL NATURAL LANGUAGE UTTERANCES USING SCFG

MOTIVATION There are in general two ways to perform behavior testing on a model: one with automatically generated data, the other with manually curated data. In this work we focus on the former because it not only scales with almost no additional cost, but also serves as a pre-filtering mechanism before we test it further with human-in-the-loop. The input to text-to-SQL models is a *natural* question. However, generating natural language has two challenges: (i) it is difficult to automatically produce novel human-like utterances with high-fidelity; (ii) natural language is inherently ambiguous, while input to text-to-SQL models is required to be accurate enough to have a one-to-one mapping between the natural question and the SQL query. Motivated by the above requirements, we propose using the inherently non-ambiguous **S**ynchronous **c**ontext-**f**ree **g**rammar (SCFG) for generating canonical natural language utterances in English[1].

DETAILS OF SCFG SCFG is a type of formal grammar which produce pairs of utterances that share a meaning with each other. There are two key components of a context-free grammar: *symbols* and *production rules* that connect them. In our case, the symbols correspond to the SQL elements, which are presented in the first column of Table 19.[2] The production rules are mappings between SQL elements and natural language words. In Figure 12 we provide such an example where SCFG maps the SQL element `DISTINCT` to the word "*unique*", hence converting the SQL query "SELECT DISTINCT Column FROM Table" to the natural language question "Select unique Column from Table". The mappings between symbols and query words are intentionally designed to mimic the language in the Spider dataset (Yu et al., 2018), which ensures that the generated examples remain close to the training distribution.[3]

Intuitively, questions produced by the SCFG lie somewhere in-between natural language and SQL: they are not as natural as real human questions, but are much more human-like than the SQL queries. Accommodating such a trade-off ensures that the generated queries are both natural and accurate. More examples of SCFG rules can be found in Appendix A.5.

GENERATION OF EVALUATION DATA To thoroughly evaluate each SQL element, we create as many valid question-query pairs as possible for each database in Spider, so that there is adequate representa-

1 This method is also extendable to other languages.

2 We collected the SQL elements from https://www.w3schools.com/sql/ and https://www.techonthenet.com/sqlite/.

3 Competent performance across categories in Figure 13 demonstrate our data overlap with the training distribution.

	Target SQL Element	#	Exact Set Match Acc. BRIDGE†	RATSQL+ RoBERTa	RATSQL+ GraPPa	RATSQL+ GAP
	Spider Dev	1034	68.2	69.6	73.4	71.8
Basic Clauses	SELECT	1700	53.6	46.5	62.6	73.5
	DISTINCT	850	86.4	86.6	94.5	88.3
	WHERE	1003	73.2	70.3	84.4	82.1
	ORDER BY	1946	51.0	54.7	71.4	76.5
	GROUP BY	653	35.5	51.3	45.9	5.7
	HAVING	604	0.1	0.0	0.0	0.0
	Cat. Avg.		53.4	53.7	**65.7**	64.4
Aggregate Ops	MIN	794	74.5	59.1	93.7	83.2
	MAX	794	75.3	17.5	85.9	47.4
	SUM	794	66.0	71.1	52.2	52.1
	COUNT	850	34.4	56.3	70.3	66.8
	AVG	794	56.7	58.1	81.8	79.7
	Cat. Avg.		61.0	52.5	**76.7**	65.9
Condition Ops	$\leqslant, <, >, \geqslant$	440	55.2	37.9	61.3	88.6
	$!=$	397	27.2	68.3	62.4	92.4
	BETWEEN	256	65.9	26.7	34.9	51.0
	Cat. Avg.		49.4	44.3	52.9	**77.3**
Logic Ops	AND	401	3.2	4.5	7.2	16.2
	OR	401	5.1	5.0	8.2	17.1
	AND & OR	369	4.1	4.3	8.6	18.1
	Cat. Avg.		4.1	4.6	8.0	**17.1**
	Overall Avg.		45.0	42.9	55.3	**55.6**

Figure 13: Results on the models per our SCFG categories. # shows the number of test examples present. Cat. Avg. reflects the category average weighted by the number of examples per each target SQL element. †BRIDGE results are averaged across three checkpoints with different random initializations, while the RATSQL results are based on the best checkpoints according to the dev set evaluation.

tion for infrequent categories. Note that many databases have tables that only correspond to a subset of elements.[4] Consequently the number of collected examples in Figure 13 (second column) are not evenly distributed.[5]

When generating examples for a given SQL element, the example operates over only one table, and we only introduce the minimum amount of other elements to make the generation grammatical

4 For example, a table with only text-type columns can not be used to generate pairs with mathematical concepts *minimum* or *less than*.

5 To have a uniform distribution, one may perform sub-sampling (which wastes valuable data), or design a model to automatically generate new tables – we leave the latter as future work.

Columns	#	Exact Set Match Acc.			
		BRIDGE	RATSQL+		
			RoBERTa	GraPPa	GAP
1	852	69.1	52.3	70.8	**85.4**
2	253	60.9	68.8	81.0	**88.9**
3	191	68.3	63.4	**85.9**	**85.9**
4	154	21.0	32.5	61.0	**81.8**
5	122	0.0	0.0	0.0	0.0
6	69	0.0	0.0	0.0	0.0

Table 16: Performance of models on SELECT clauses by number of columns being selected.

and uncompounded. For example, the operator `BETWEEN` necessitates `SELECT` and `WHERE` clauses to generate a coherent query, but any additional operators, even if they can make the query more compositional, are excluded, as our goal is to unit test each SQL element individually. In turn, our generated data are also intended to be as easy as possible for models to succeed on.

	Target SQL Element and Example	Model Predictions with Highlighted Errors
SELECT	***NL:*** Select name, id, department name, total credits from student ***SQL:***`SELECT name, ID, dept_name, tot_cred FROM student`	***BRIDGE:***`SELECT student.ID, student.name, student.dept_name, student.tot_cred FROM student` ***RS+RoB:***`SELECT student.name, student.ID, student.dept_name, Sum(student.tot_cred) FROM student GROUP BY student.ID` ***RS+GraPPa:***`SELECT student.name, student.ID, student.dept_name, Sum(student.tot_cred) FROM student` ***RS+GAP:***`SELECT student.name, student.ID, student.dept_name, Sum(student.tot_cred) FROM student`

Figure 14: Model predictions on a randomly chosen `SELECT` example. See Appendix A.5 for additional qualitative examples of model predictions on different categories.

HUMAN VERIFICATION OF EVALUATION DATA To verify that our generated examples are indeed human-like and accurate, we recruited volunteers[6] who are proficient in SQL to label a subset of 40 randomly chosen question-query pairs, and rate each pair on its "readability" and "semantic equality". The question-query pairs are chosen such that all categories are represented at least twice. Each question-query pair was annotated by three annotators and we take their majority vote. An example given to annotators can be found in the Appendix A.4.

6 Our annotation task posed no risk or harm to annotators, and required 30 minutes of the volunteers' time.

6.5 EXPERIMENTS

6.5.1 *Experiment Setup*

MODELS We evaluate four leading models on the Spider challenge (Yu et al., 2018) on our generated question-query pairs: BRIDGE (Lin, Socher, and Xiong, 2020), RATSQL-RoBERTa and RATSQL-GraPPa (Yu et al., 2020) and RATSQL-GAP (Shi et al., 2020). With the exception of BRIDGE, the other models were developed upon the original RATSQL model (Wang et al., 2020), which was notable for introducing a *relation-aware self-attention mechanism* for schema linking. Yu et al. (2020) extended the RATSQL framework by adding pre-training into their setup, and Shi et al. (2020) also incorporates supplementary pre-training triplet data generated by another model. The BRIDGE model is fundamentally different from the others, as it consists of a sequentially-driven architecture, rather than operating over graphs. For schema-linking, BRIDGE uses a custom encoder powered by BERT (Devlin et al., 2019) with attention over the sequences.

EVALUATION METHODOLOGY Our experiments consist of evaluating each model on the generated set of question-query pairs with the canonical language questions as inputs. We evaluate Exact Set Match Accuracy for subsets of the data pertaining to each target SQL element, and then calculate the average score for each SQL token category weighted by number of examples.

6.5.2 *Results*

MAIN RESULTS Figure 13 highlights several interesting observations.[7] Most models only perform on par with their baseline (or better) on a few target SQL elements (e.g. `DISTINCT`, `WHERE`). More often they perform below the baseline on most elements, with a few extreme outliers for total or near total failure (e.g. `HAVING`, `AND`).

CONTROLLED EVALUATION All models perform below their own baseline accuracies for simple examples that test the `SELECT` clause. We present an example of such model predictions in Figure 14. One contributing factor to these low scores is the number of columns being selected. Table 16 shows that SQL models are only able to successfully produce queries with a limited number of columns, although basic column selection should not be such a difficult task for these models. While it is not surprising that models show difficulty generalizing to unseen length or structures (Lake and Baroni, 2017), this finding

7 The metrics in Figure 13 are diagnostic instead of explanatory. There can be multiple factors affecting the model performance on an evaluation point and our tests cannot isolate them.

is concerning because there are many practical use cases where users will need to select more than four columns.[8]

6.6 CONCLUSION

We propose a simple and controllable approach for synthesizing text-to-SQL pairs for unit testing model performance on various semantic categories. Our controlled test suites allow for more extensive and fine-grained evaluation of state-of-the-art text-to-SQL models, which reveal a general lack of robustness in generalizing beyond the benchmark examples across several categories such as `SELECT` and `WHERE`. More importantly, our study highlights the importance of developing evaluation strategies beyond fixed test and dev set accuracy for understanding real progress made by the state-of-the-art text-to-SQL models and the remaining key challenges.

8 For example, the large tables in Spider's *soccer_1* database

7

COMMON SENSE BIAS IN SEMANTIC ROLE LABELING

7.1 ABSTRACT

Large-scale language models such as ELMo and BERT have pushed the horizon of what is possible in semantic role labeling (SRL), solving the out-of-vocabulary problem and enabling end-to-end systems, but they have also introduced significant biases. We evaluate three SRL parsers on very simple transitive sentences with verbs usually associated with animate subjects and objects, such as *Mary babysat Tom*: a state-of-the-art parser based on BERT, an older parser based on GloVe, and an even older parser from before the days of word embeddings. When arguments are word forms predominantly used as person names, aligning with common sense expectations of animacy, the BERT-based parser is unsurprisingly superior; yet, with abstract or random nouns, the opposite picture emerges. We refer to this as *common sense bias* and present a challenge dataset for evaluating the extent to which parsers are sensitive to such a bias. Our code and challenge dataset are available here: `github.com/coastalcph/comte`

7.2 INTRODUCTION

Semantic role labeling (SRL) refers to a shallow semantic dependency parsing that returns predicate-argument structures for input sentences; see Figure 15. Modern-day SRL systems, like most other NLP technologies, rely heavily on large-scale language models. Such language models are extremely useful for generalizing to out-of-vocabulary items, making subtle syntactic distinctions, and for capturing a range of lexical ambiguities; but they also introduce notable biases.

Previous work has shown that SRL systems exhibit demographic biases (Zhao et al., 2017); we focus on a form of belief bias (Sternberg and Leighton, 2004), which we will refer to as *common sense bias*, reflecting how language models encode conventional associations, which in many ways are indistinguishable from common sense (Trinh and Le, 2019). While demographic biases can lead to discrimination against under-represented demographics, belief biases can lead to discrimination against rare events; or, more precisely, lead SRL systems to err on sentences that express unlikely states of affairs. This is what belief biases refer to in cognitive science (Sternberg and Leighton, 2004): human preferences for conclusions that align with values, beliefs, and prior knowledge. Belief biases in models can, like

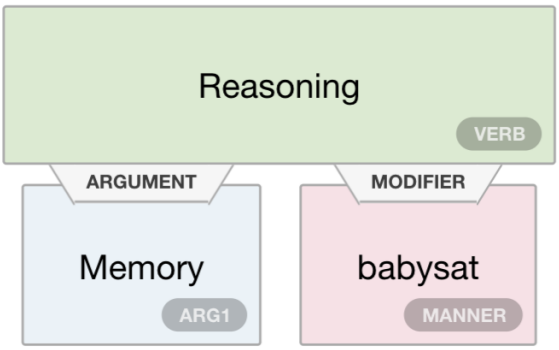


Figure 15: The (incorrect) analysis of *Memory babysat Reasoning* by Shi and Lin (2019).

demographic biases, exacerbate societal challenges, e.g., anomaly detection, and also correlate with demographics, since groups differ in how much they engage with counterfactual and fictitious contents.

We compare the errors of a modern, competitive SRL system (Shi and Lin, 2019), based on BERT (Devlin et al., 2019), and show how it, unlike earlier SRL systems, suffers from common sense bias: When confronted with sentences that, when read literally, express unlikely states of affairs, it can ignore obvious cues and produce false predicate-argument structures even for very simple sentences. The sentence in Figure 15, for example, can be understood as expressing that the abstract concept of *Memory* babysat the abstract concept of *Reasoning*. The literal reading represents an unlikely state of affairs, since abstract concepts generally do not have the capacity of babysitting. Obviously, this does not prevent language users from uttering the sentence, and it is, for most of us, not hard to make sense of it: The sentence, for example, could mean something like *memory assists reasoning*. Many similar sentences can be found in the wild, e.g., *the US babysits Israel* (from `cnn.com`) or *Love bodyslams you* (from `quizlet.com`). Other sentences express unlikely states of affairs, not because of linguistic creativity, but because they refer to possible worlds, not ours, for scientific, literary, political or other reasons. We believe it is critical that SRL parsers should be robust to such variation, but our experiments show that while SRL performance numbers have gone up dramatically in recent years, parsers seem to have become more sensitive to it.

CONTRIBUTIONS We present an error analysis of three very different SRL parsers for English: the supervised, log-linear, quadratic-time parser proposed in Björkelund, Hafdell, and Nugues (2009); the supervised, deep, linear-time parser proposed in Stanovsky et al. (2018), based on GloVe embeddings (Pennington, Socher, and Manning, 2014) and recurrent networks (Hochreiter and Schmidhuber, 1997); and the self-supervised (and supervised), deep, linear-time parser proposed

names	Mary	babysat	Tom
countries	Sweden		Senegal
abstract	Memory		Reasoning
random-nouns	apples		oranges
random-strings	asdfabv		sadlfkj

Figure 16: Examples of transitive sentences with person names, country names, abstract nouns, (randomly chosen) plural common nouns, or random strings as arguments. Person names, and to some degree country names (which are often personified (Wang, 2020)), align with expectations of animacy.

Ref	Model	F_1
Björkelund, Hafdell, and Nugues (2009)	MST/MIRA	0.803
Stanovsky et al. (2018)	LSTM/GloVe	0.823
Shi and Lin (2019)	BERT	0.888

Table 17: The three SRL systems used below and their performance on the CoNLL 2005 benchmark

in Shi and Lin (2019), based on BERT (Devlin et al., 2019). Instead of evaluating these models on standard benchmarks of newspapers, where predicate-argument structures already align with the 'beliefs' of BERT, we evaluate the systems on randomly generated transitive sentences of the form NP-V-NP, with V expressed by verbs strongly associated with A_0-V-A_1 frames, and the NPs expressed by proper nouns, abstract nouns or plural common nouns. From these experiments, we show that (a) the SRL systems considered here frequently err on such sentences; (b) the SRL error distribution across verb lemmata is uncorrelated with the errors of a dependency parser; (c) what pairs of NP semantic categories lead to errors for what verbs; and (d) how the BERT-based system suffers from common sense bias. Finally, we create a 1000-sentence challenge dataset for probing SRL for common sense bias. Our error analyses paint a complementary, yet entirely different picture of what SRL systems can and cannot, compared to previous work (He et al., 2017; Strubell et al., 2018), which has focused on long-distance dependencies and the need for syntax.

7.3 SEMANTIC ROLE LABELING SYSTEMS

Björkelund, Hafdell, and Nugues (2009) combine three logistic regression classifiers with beam search and a global reranker: the first classifier identifies predicates, the second their arguments, and the third labels the semantic dependencies between predicates and their arguments. The system relies on a POS tagger and a syntactic dependency parser to generate features for the classifiers. This system had the second-best performance in the CoNLL 2009 Shared Task. Stanovsky

Verb	Error	A_0VA_2	A_1VA_2	...V	Expl
fails	0.898	0.006	**0.758**	0.000	Syntax
calls	0.528	**0.467**	0.020	0.020	
trips	0.356	0.007	0.000	**0.128**	POS
tips	0.875	0.010	0.010	**0.687**	
bodyslams	0.373	0.065	0.052	0.034	?
babysits	0.212	0.048	0.072	0.035	

Table 18: Error rates and most frequent error types for common verbs in their present and past tense forms, in simple SOV constructions, e.g., *John calls Mary*. All numbers are for Shi and Lin (2019). Boldfaced error types most frequent (of the four presented here). The verbs *bodyslams* and *babysits* are used in our experiments, because (a) they have strong selectional restrictions for animate subjects and objects, (b) they predominantly realize A_0 and A_1 as subjects and objects (unlike *fails* and *calls*), and (c) while all English verbs tend to have noun readings, the verb readings are far more frequent (unlike for *trips* and *tips*).

et al. (2018) rely on a standard recurrent architecture. They use GloVe embeddings (Pennington, Socher, and Manning, 2014), in conjunction with embeddings from a POS tagger, and stack bidirectional LSTM layers (Hochreiter and Schmidhuber, 1997) on top of the embedding layer. The representation at each time-step is passed to a classifier, which directly predicts the output label for that time-step. Unlike Björkelund, Hafdell, and Nugues (2009), they do not rely on search over possible output combinations. Shi and Lin (2019) also do not rely on search, but reduce SRL to two-stage sequence labeling, both stages pretrained with BERT-large (Devlin et al., 2019); first identifying predicates, then arguments, while conditioning on the predicates.

7.4 COARSE-GRAINED ERROR ANALYSIS

In our error analysis, we focus on simple three-word sentences that consist of a noun, a transitive verb, and a noun. The transitive verbs are hand-picked to exhibit strong preferences for animate subjects and objects, low ambiguity, and predominantly realize their agents (A_0) as subjects, and their second argument (A_1) as objects. The error analysis consists of comparing performance across different types of subjects and objects and comprises examples such as those in Figure 16. The arguments exhibit various degrees of animacy associations, aligning more or less with common sense expectations. We obtain the names from the NAMES library,[1] the country names from

1 https://pypi.org/project/names/

Error	Björkelund, Hafdell, and Nugues (2009)	Stanovsky et al. (2018)	Shi and Lin (2019)
names	0.158	**0.341**	0.077
countries	0.183	**0.505**	0.030
abstract	0.174	**0.353**	0.133
random-nouns	0.188	0.287	**0.310**
random-strings	**0.997**	0.172	0.313

Table 19: **Main results:** Error rates of three SRL systems across transitive sentences with person names in subject and object positions, versus country names, abstract nouns, (randomly chosen) plural common nouns, or random strings in those positions

WorldMap,[2] the abstract nouns from YourDictionary,[3] and common nouns from the Princeton WordNet.[4]

We assume a correct semantic parse associates subject with A_0 and object with A_1 (of the predicate introduced by the verb). This is obviously not true for all verbs (Hovy et al., 2006b; Palmer, Gildea, and Kingsbury, 2005). In Table 18, we list verbs that frequently associate subjects and objects with other arguments (*fails* and *calls*), as well as verbs that are very ambiguous and easily mistaken for nouns (*trips* and *tips*). Both phenomena are reflected in the distribution of analyses for Shi and Lin (2019). While much can be said about these verbs, our main contribution here is highlighting the role of common sense bias in some SRL parsers, and we thus focus on verbs where we can safely assume a A_0VA_1 reading is correct (such as *babyslams* and *babysits*).[5]

Error analysis results are presented in Table 19. If performance drops considerably below the performance label with names or countries, when using abstract nouns, randomly sampled nouns, or sim-

2 http://worldmap.harvard.edu/

3 https://examples.yourdictionary.com/examples-of-abstract-nouns.html

4 https://wordnet.princeton.edu/

5 The six verb lemmata we use are: *bodyslam*, *bodypaint*, *comb*, *manicure*, *elbow*, and *babysit*.

Tajikistan bodyslams Maldives	Lebanon bodyslammed Netherlands
Myanmar bodyslammed Andorra	Bangladesh bodypaints Peru
Luxembourg bodypainted Andorra	Kazakhstan bodypainted Guinea
Bangladesh combed Turkey	Bangladesh manicured Swaziland

Table 20: Simple sentences on which Stanovsky et al. (2018) and Shi and Lin (2019) both err. Björkelund, Hafdell, and Nugues (2009), in contrast, assigns correct parses to all of these. Try yourself: barbar.cs.lth.se:8081/

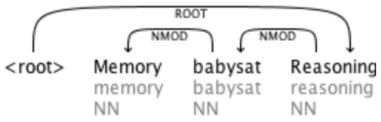


Figure 17: Parse tree in Björkelund, Hafdell, and Nugues (2009) for *Memory babysat Reasoning*.

ply random strings, as arguments, this suggests a common sense bias, seen very strongly with Shi and Lin (2019). Björkelund, Hafdell, and Nugues (2009), in contrast, exhibits near-uniform performance across the different sets of arguments. Since the parser has no strategy to deal with out-of-vocabulary items, it exhibits worse performance on random strings.[6] Stanovsky et al. (2018), surprisingly, seems extremely sensitive to country name arguments,[7] and performance oddly improves with random strings arguments. Since these are out-of-vocabulary, the parser probably drops back to a default strategy. Notably, Shi and Lin (2019) does well on country names, there are plenty of examples that Stanovsky et al. (2018) and Shi and Lin (2019) get wrong, but that Björkelund, Hafdell, and Nugues (2009) get right; see Table 20 for examples.

COMPARISON WITH DEPENDENCY PARSER ERRORS While only one of the parsers (Björkelund, Hafdell, and Nugues, 2009) relies on input features from a syntactic parser, it is tempting to think that, in line with previous error analyses of SRL systems (He et al., 2017; Strubell et al., 2018), the error distribution can be explained by syntactic ambiguities and resulting syntactic errors. This, perhaps unsurprisingly, turns out to reliably explain the error distribution observed with Björkelund, Hafdell, and Nugues (2009). See Figure 17 for the syntactic parse on *Memory babysits Reasoning*, on which the log-linear parser fails to deliver any SRL analysis, interpreting the three-word sentence as a nominal compound. For the neural parsers, there is no correlation, however. We ran a syntactic parser (Dozat, Qi, and Man-

6 Björkelund, Hafdell, and Nugues (2009) near-consistently analyze these as intransitive with the first two words making up A_1.

7 We found no explanation for Stanovsky et al. (2018)'s poor performance with country name arguments.

ning, 2017) on our three-word sentences and correlated errors across lemmata. We observed a small, but *negative* correlation between error rates.

7.5 FINE-GRAINED ERROR ANALYSIS

MULTITUDE OF ERRORS Our first observation is that across all verb lemmata, the parser in Shi and Lin (2019) produces *many* different output trees, depending on the argument word forms. For some lemmata, the error distribution is near-uniform across 15-20 outputs. It is well-established in SRL that infrequent contexts lead to low confidence (Chen, Palmer, and Sporleder, 2011), explaining why common sense bias leads to a multitude of errors.

MORPHOSYNTACTIC AMBIGUITY While parsing errors do not correlate with errors of Shi and Lin (2019) (§7.4), the SRL system seems to be sensitive to part-of-speech ambiguity. It errs, for example, on *Insomnia trips jaywalking*, but not on *Insomnia tripped jaywalking*, presumably because *trips* is (on its own) ambiguous.[8] Sensitivity to such ambiguities disappears when aligning with common sense: The parser does not err on *Mary trips John* or *Mary likes jaywalking*. The same ambiguity leads to error in *London trips John.*, but not in *London tripped John.* With the even more frequent surname of *Washington*, the effect disappears, and Shi and Lin (2019) get both verb forms right.

7.6 COMTE: A TEST OF COMMON SENSE BIAS

Our challenge dataset[9] COMTE consists of 1,000 simple, three-word sentences with the same gold analysis: the second word is the predicate, the first word its A_0, the last word its A_1. The predicates are sampled at random from a list of six carefully selected verbs (see §3) that select for animate subjects and objects and consistently prefer these to be A_0 and A_1. As before, we combine the verbs with names, countries, abstract nouns, plural common nouns, and random strings. The sentences were simply the first 1,000 sentences that we sampled this way, with 200 sentences in each category (names, countries, etc.) – and which satisfied a simple criterion: Neither Shi and Lin (2019) nor (Stanovsky et al., 2018) would get it right. COMTE, in other words, consists of 1,000 trivial sentences that two competitive SRL parsers failed to parse correctly.

What can COMTE be used for? Obviously, it can not be used to fine-tune parsers on, for example. It would take only a few examples to

8 This is orthogonal to the ambiguity of *jaywalking*; see Padó, Pennacchiotti, and Sporleder (2008) for the analysis of nominal predicates.

9 Our dataset differs from previous challenge datasets for mixed language (Pal and Sharma, 2019), chat (Rachman et al., 2018), etc., in being synthetic.

learn what is going on in the data, and training would likely lead to over-fitting. COMTE can also not be used to derive parsing performance figures that tell us much about the performance of parsers in the wild. The 1,000 sentences should, in our view, be thought of as a single probe into the degree to which a parser is sensitive to common sense bias. A parser should rarely err on the examples in the challenge dataset: They are all trivially simple, and while some argument words can be ambiguous, the verbs so strongly select for simple A_0VA_1 frames that parsers should unambiguously prefer this reading. If they don't, this is a sign they struggle with simple transitive sentences, like Stanovsky et al. (2018), or that they are prone to common sense bias, like Shi and Lin (2019). In order to quantify the degree to which the effect can be attributed to common sense bias, performance with *names* can be used as a baseline: If performance is much better for names than for some of the other categories, like with Shi and Lin (2019), this is an indicator of common sense bias.

Part IV

APPENDIX

APPENDIX

A.1 FULL RESULTS (CHAPTER 2)

Singlish								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	46.77	68.89	74.34	41.07	42.89	66.76	74.17
	DRO-One	44.23	64.73	71.90	49.18	40.73	63.05	70.13
	DRO-Random	43.33	65.63	71.58	49.14	39.07	61.02	68.42
	DRO-Language	43.19	64.80	71.22	48.88	39.57	61.54	70.34
mBERT	ERM	47.78	68.82	75.77	42.26	37.00	61.71	70.85
	DRO-One	44.37	65.13	72.54	50.99	33.71	57.79	65.72
	DRO-Random	44.34	65.95	72.44	50.39	35.25	59.20	66.93
	DRO-Language	43.69	64.91	71.61	50.49	33.45	58.91	67.42

Table 21: Full results for Singlish Mixed-Language experiments.

Naija								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	63.83	80.52	85.44	42.41	59.97	78.72	83.93
	DRO-One	60.99	77.52	82.94	52.51	56.76	76.21	81.94
	DRO-Random	60.40	78.44	82.88	52.69	56.33	75.27	81.17
	DRO-Language	60.40	77.40	82.69	54.18	54.80	74.48	80.32
mBERT	ERM	62.68	80.52	85.55	44.98	62.19	82.25	87.08
	DRO-One	60.76	77.74	82.88	58.15	56.43	77.61	82.67
	DRO-Random	60.34	77.23	82.24	57.90	56.70	77.50	82.51
	DRO-Language	58.88	76.56	81.73	59.91	54.99	76.63	82.50

Table 22: Full results for Nigerian Pidgin Mixed-Language experiments.

Haitian								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	68.09	82.98	87.34	55.05	43.35	63.89	71.35
	DRO-One	57.04	71.12	75.58	121.51	36.73	52.55	58.25
	DRO-Random	57.65	71.53	75.79	119.17	36.16	50.63	56.38
	DRO-Language	57.55	71.23	75.28	118.85	36.69	50.48	55.89
mBERT	ERM	60.79	76.70	81.56	60.27	46.35	64.56	70.96
	DRO-One	51.06	65.45	69.71	148.12	34.57	49.30	55.61
	DRO-Random	50.86	65.05	69.50	146.18	34.52	49.58	55.00
	DRO-Language	50.15	64.54	69.40	145.97	33.55	48.21	55.08

Table 23: Full results for Haitian Mixed-Language experiments.

Singlish								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	53.80	75.02	80.36	34.22	51.26	74.09	80.15
	DRO-One	45.34	64.41	70.14	66.53	43.59	63.42	69.33
	DRO-Random	45.73	64.66	71.00	64.16	42.40	64.38	70.74
	DRO-Language	44.73	65.16	71.08	57.54	40.57	62.68	69.78
mBERT	ERM	56.81	77.03	81.65	34.49	46.87	72.55	79.49
	DRO-One	47.49	65.84	70.97	76.57	36.17	56.74	64.51
	DRO-Random	47.85	65.88	70.93	74.87	37.66	58.37	65.41
	DRO-Language	45.77	64.77	70.39	68.55	33.94	55.01	62.09

Table 24: Full results for Singlish Creole-Only experiments.

Naija								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	73.72	88.62	91.99	28.14	71.38	87.33	90.94
	DRO-One	64.28	79.37	83.95	61.81	59.86	77.00	81.60
	DRO-Random	63.72	79.57	83.92	60.31	59.31	75.55	80.29
	DRO-Language	63.58	79.48	84.29	56.83	59.74	77.08	81.84
mBERT	ERM	72.96	87.58	91.15	31.77	70.42	87.58	91.42
	DRO-One	63.72	78.36	82.77	76.24	60.78	77.07	81.78
	DRO-Random	63.52	77.77	82.18	74.53	61.02	78.01	82.88
	DRO-Language	63.13	78.16	82.80	71.77	60.73	77.37	82.25

Table 25: Full results for Nigerian Pidgin Creole-Only experiments.

Haitian								
Model	Strategy	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
BERT	ERM	73.15	86.12	88.55	55.51	55.50	71.76	77.94
	DRO-One	58.16	71.23	75.48	144.47	36.91	51.39	56.04
	DRO-Random	57.65	70.52	75.38	142.04	37.41	52.55	57.72
	DRO-Language	56.94	71.33	74.97	138.60	35.50	49.66	55.03
mBERT	ERM	66.06	80.45	84.30	69.25	55.58	72.49	78.72
	DRO-One	50.35	65.05	69.10	174.45	35.86	51.60	56.72
	DRO-Random	48.63	64.03	67.78	172.26	32.54	48.31	53.60
	DRO-Language	49.14	64.24	68.69	167.90	34.59	49.34	55.45

Table 26: Full results for Haitian Creole-Only experiments.

Dataset	Model	P@1	P@5	P@10	PLL	P_D@1	P_D@5	P_D@10
Singlish	BERT	23.94	38.49	45.09	76.01	21.09	36.65	42.22
	mBERT	14.30	23.12	27.03	92.97	10.23	23.57	29.85
Nigerian Pidgin	BERT	22.79	34.04	39.88	142.66	10.92	18.07	22.96
	mBERT	14.90	26.34	31.87	153.54	8.08	16.24	20.72
Haitian Creole	BERT	18.84	30.60	37.59	177.40	5.65	11.89	16.29
	mBERT	11.96	22.39	27.96	175.14	7.10	12.20	16.76

Table 27: Full results for pretrained baselines.

A.2 FULL RESULTS (CHAPTER 3)

A.2.1 *Training Setup*

Type	Target	Training Langs	Train Size (#Sents)
Creole	acf	fra, hau, yor, ibo	38,140
	hat	fra, fon, ibo, spa	31,669
	jam	eng, hau, spa, ibo	44,545
	pcm	eng, hau, yor, por	35,189
Non-Creole	dan	nno, isl, swe, deu	39,354
	spa	fra, por, ita, rom	30,870
Control	-	afr, chr, hun, quy	37,398

Table 28: Details of the data used for training our experiments. The same dataset was used to train "Control" experiments, for every Target language in this table. For the Train Size, the #sents is determined by taking the parallel bible verses for each of the Training Lang(uage)s, and using a sentence splitter to obtain the training examples. All experiments had a Dev Size of 500 bible verses ($\approx$ 500 sentences), for all languages (Target+Training).

A.2.2 *Results*

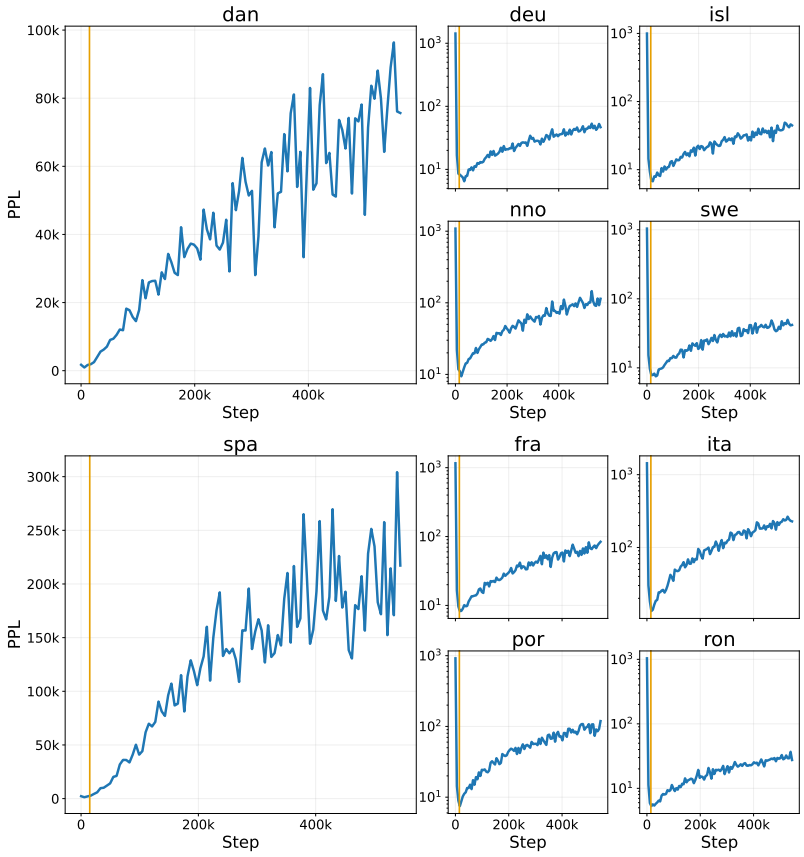


Figure 18: Full results for zero-shot transfer to non-Creole languages when training on their related languages. Before 100 epochs (shown at the yellow line), perplexity drops for the non-Creoles, as expected. As the model overfits to the training languages over time, perplexity climbs steadily.

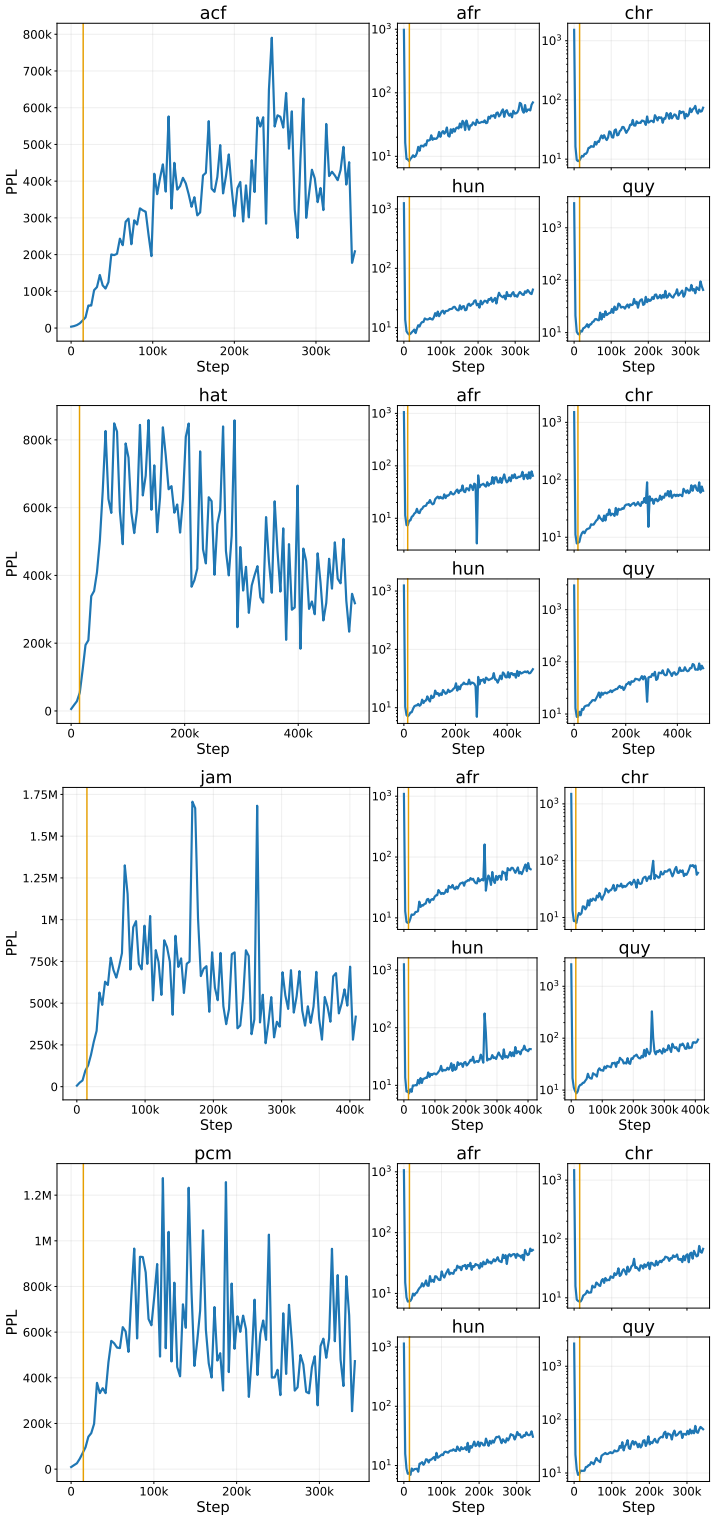


Figure 19: Full results for zero-shot transfer for Creole languages when training on random languages. The yellow line marks 100 epochs of training. Although the training languages are not related to the Creoles, we still observe the two-phase pattern, in which perplexity for Creoles drops after overfitting.

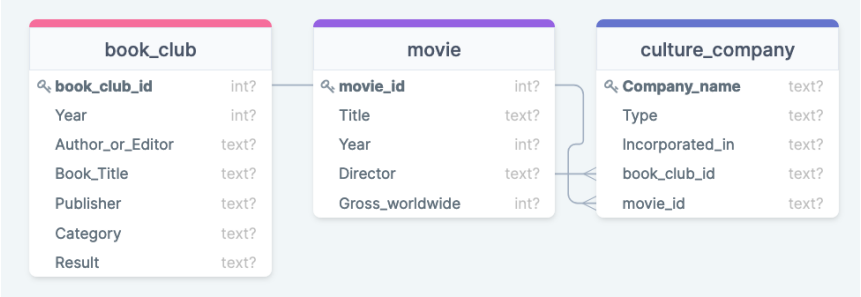
A.3 MODEL PERFORMANCE ON DEV EXAMPLES CORRESPONDING TO CATEGORIES (CHAPTER 6)

Target SQL Element	#train	#dev	Exact Set Match Acc. BRIDGE	RATSQL+ RoBERTa	RATSQL+ GraPPa	RATSQL+ GAP
SELECT	213	32	82.3	90.6	96.9	81.2
DISTINCT	113	5	86.7	100	100	60.0
WHERE	343	61	77.2	83.6	83.6	100
ORDER BY	560	83	78.3	88.0	90.4	78.3
GROUP BY	16	8	83.3	50.0	50.0	75.0
MIN	2	4	16.7	0.0	50.0	0.0
MAX	10	5	0.0	0.0	0.0	0.0
SUM	25	2	100	100	100	100
COUNT	245	40	99.2	100	97.5	97.5
<=, <, >, >=	70	6	77.8	66.7	100	100
!=	14	52	83.3	85.7	85.7	100
AND	50	5	66.7	60.0	100	60.0
OR	54	10	100	88.0	100	78.3

Figure 20: Performance of models on Spider Dev by our categories. SCFG elements that had zero corresponding examples are removed from the table. Here we include the number of examples in Spider training and Spider dev to demonstrate the underlying training and development distributions. Examples counted here are strictly relate to the chosen category. (i.e. examples with multiple SQL elements that do not pertain exactly to the categories are excluded from these counts).

A.4 EXAMPLE OF ANNOTATION TASK (CHAPTER 6)

Example database schema given to annotators:



Example question-query pair given to annotators:
/**Question:** Select year from movie when movie id is greater than 1
Query: `SELECT Year FROM movie WHERE movie_id > 1 ;`

Annotators are asked to choose one answer from the list below, to describe the readability and equivalency of the question-query pair, above:

1. **Readability**:
   - I can easily understand the question
   - I have some problems understanding the question, but I can understand with some effort
   - I do not understand the question after trying my best to interpret it
2. **Equivalency**:
   - The question and the SQL query match perfectly
   - The question and SQL query do not fully match, but the answer to the question can be inferred from the SQL query results
   - The SQL query does not return the answer to the question

		#Votes
(R)	Easily understandable	94
	Understandable with effort	22
	Not understandable	4
(E)	Perfect match	114
	Question inferred from SQL	2
	Query does not return answer	4

Table 29: Full annotation results for **R**eadability and **E**quivalency.

For both annotation tasks, the same 4 pairs were chosen as bad:

	Pairs chosen by annotators as Not-Readable and Not-Equivalent	**Problem**
1	"Select unique date contact to from organization contact individuals" `SELECT DISTINCT date_contact_to FROM Organization_Contact_Individuals;`	awkward column name
2	"Select the number of sequence length from protein" `SELECT COUNT(sequence_length) FROM protein;`	column missing from schema image
3	"Select number city affected from affected region when storm id equals 1" `SELECT Number_city_affected FROM affected_region WHERE Storm_ID = 1;`	awkward column name
4	"Select the average value of launch from program" `SELECT AVG(Launch) FROM program;`	column missing from schema image

A.5 EXAMPLE MODEL PREDICTIONS AND SCFG PRODUCTION RULES (CHAPTER 6)

See Table 21 for model predictions, and Figures 22 and 23 for example SCFG production rules.

	Target SQL Element and Example	Model Predictions with Highlighted Errors & Highlighted Missing Elements
GROUP BY	***NL:*** Select time slot id and minimum end minute from time slot grouped by time slot id ***SQL:***`SELECT time_slot_id, MIN(end_min) FROM time_slot GROUP BY time_slot_id`	***BRIDGE:***`SELECT time_slot.time_slot_id, MIN(time_slot.end_hr) FROM time_slot GROUP BY time_slot.time_slot_id` ***RS+RoB:***`SELECT time_slot.time_slot_id, MIN(time_slot.end_min) FROM time_slot GROUP BY time_slot.time_slot_id` ***RS+GraPPa:***`SELECT time_slot.time_slot_id, Min(time_slot.end_min) FROM time_slot GROUP BY time_slot.time_slot_id` ***RS+GAP:***`SELECT time_slot.time_slot_id, Min(time_slot.end_min), Min(time_slot.end_min) FROM time_slot GROUP BY time_slot.time_slot_id`
HAVING	***NL:*** Select name from people grouped by name with maximum people id equal to 2 ***SQL:***`SELECT Name FROM people GROUP BY Name HAVING MAX(People_ID) = 2`	***BRIDGE:***`SELECT people.Name, MAX(people.Height) FROM people JOIN perpetrator ON people.People_ID = perpetrator.People_ID WHERE people.People_ID = 2 GROUP BY people.People_ID people.Name HAVING MAX(people.People_ID) =2)` ***RS+RoB:***`SELECT people.Name FROM people GROUP BY people.Name HAVING Count(*) >= MAX(people.People_ID =) 'terminal'` ***RS+GraPPa:***`SELECT people.Name, Max(people.People_ID) FROM perpetrator JOIN people ON perpetrator.People_ID = people.People_ID GROUP BY people.Name HAVING Count(*) >= MAX(people.People_ID) = 'terminal'` ***RS+GAP:***`SELECT people.Name FROM people JOIN perpetrator WHERE GROUP BY Name HAVING MAX people.People_ID = 'terminal'`
OR	***NL:*** Select building from classroom when building equals Stabler or capacity equals 115 or room number equals 261 ***SQL:***`SELECT building FROM classroom WHERE building = 'Stabler' OR capacity = 115 OR room_numer = 261`	***BRIDGE:***`SELECT classroom.building FROM classroom WHERE classroom.building = 'Stabler' OR classroom.capacity = 115 UNION SELECT * FROM classroom WHERE classroom.building = 'Stabler' OR classroom.room_numer = 261` ***RS+RoB:***`SELECT classroom.building FROM classroom WHERE classroom.building = 'terminal' OR classroom.capacity = 'terminal' OR classroom.room_number = 261` ***RS+GraPPa:***`SELECT classroom.building FROM classroom WHERE classroom.building = 'terminal' OR classroom.capacity = 'terminal' OR classroom.room_numer = 261` ***RS+GAP:***`SELECT classroom.building FROM classroom WHERE classroom.building = 'terminal' OR classroom.capacity = 'terminal' AND OR classroom.room_number = 'terminal' AND classroom.room_number = 'terminal'`

Figure 21: Example predictions on selected target SQL elements from the BRIDGE, and RATSQL (RS) based models using RoBERTa (+RoB), GraPPa, and GAP.

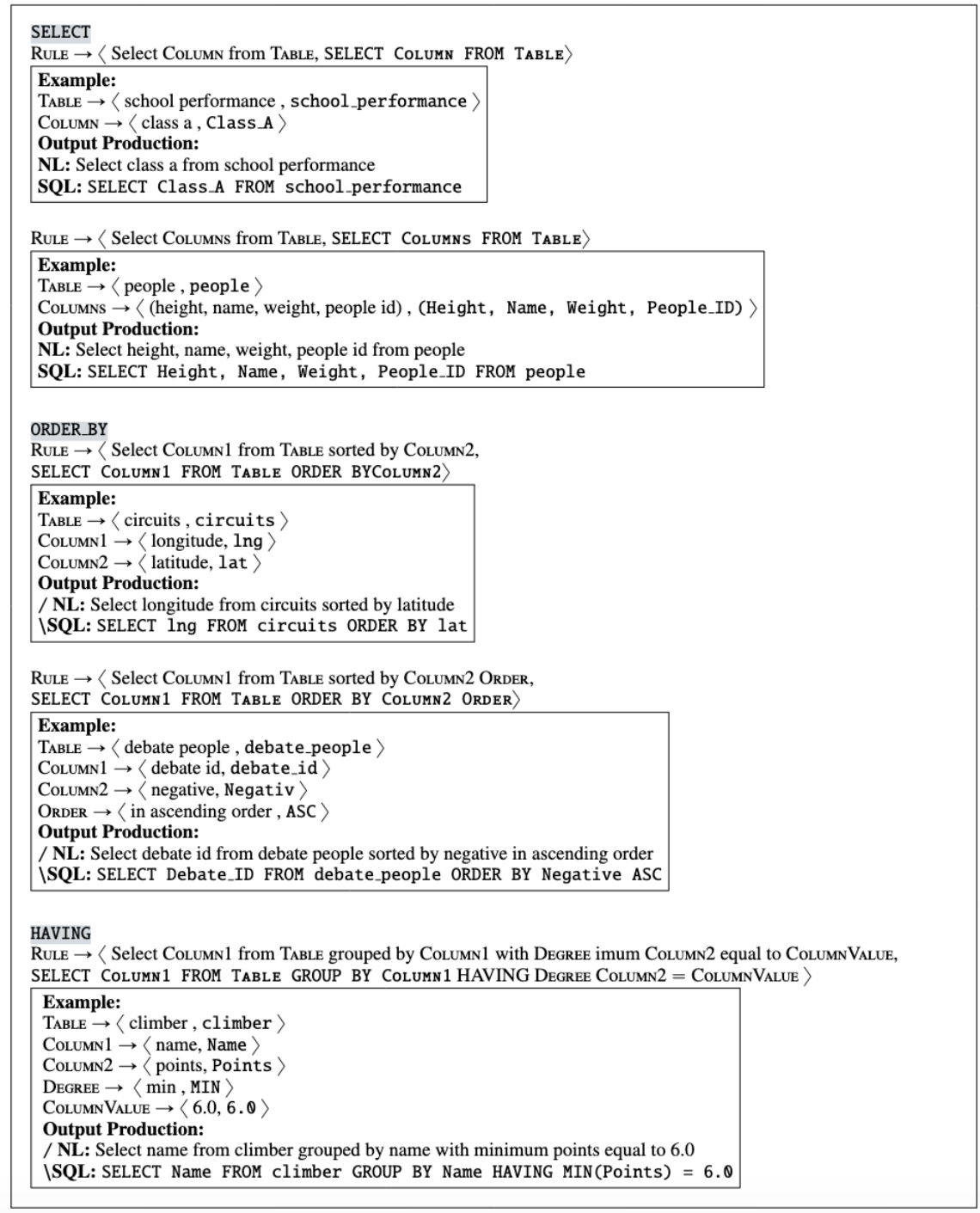


Figure 22: Example SCFG Production Rules for selected SQL Clauses

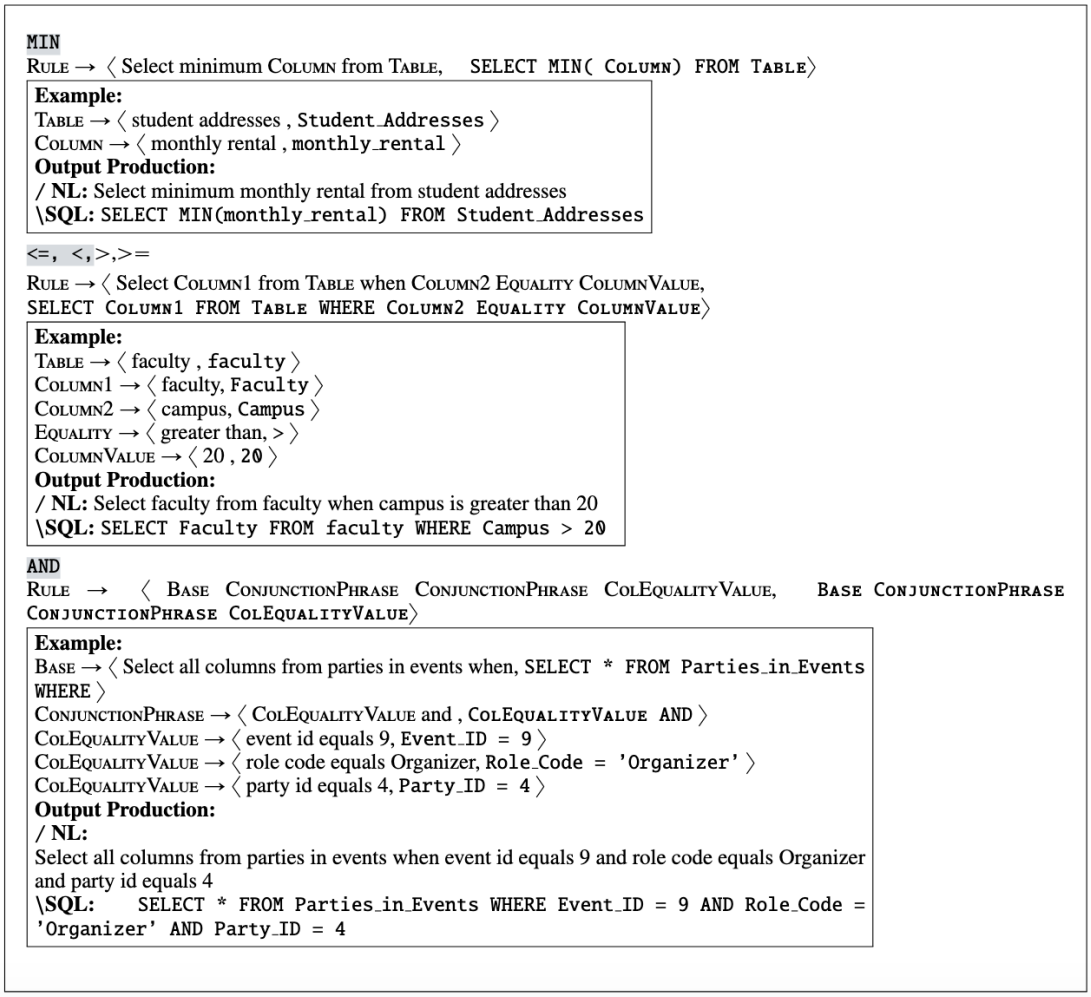


Figure 23: Example SCFG Production rules for other selected SQL operators

BIBLIOGRAPHY

Aboh, Enoch Oladé (2015). *The emergence of hybrid grammars : language contact and change*. eng. Cambridge approaches to language contact. Hastings, England: Cambridge University Press. ISBN: 0-521-15022-1.

Aboh, Enoch Oladé and Michel DeGraff (2016). "A Null Theory of Creole Formation Based on Universal Grammar." In.

Adelani, David Ifeoluwa et al. (2021). "MasakhaNER: Named Entity Recognition for African Languages." In: *Transactions of the Association for Computational Linguistics* 9, pp. 1116–1131. DOI: 10.1162/tacl_a_00416. URL: https://aclanthology.org/2021.tacl-1.66.

Agić, Željko, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard (2016). "Multilingual Projection for Parsing Truly Low-Resource Languages." In: *Transactions of the Association for Computational Linguistics* 4, pp. 301–312. DOI: 10.1162/tacl_a_00100. URL: https://aclanthology.org/Q16-1022.

Agić, Željko and Ivan Vulić (July 2019). "JW300: A Wide-Coverage Parallel Corpus for Low-Resource Languages." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3204–3210. DOI: 10.18653/v1/P19-1310. URL: https://www.aclweb.org/anthology/P19-1310.

Ahia, Orevaoghene and Kelechi Ogueji (2020). "Towards Supervised and Unsupervised Neural Machine Translation Baselines for Nigerian Pidgin." In: *ArXiv* abs/2003.12660.

Ajisafe, Daniel, Oluwabukola Grace Adegboro, Esther Oduntan, and Tayo Oladiran Arulogun (2020). "Towards End-to-End Training of Automatic Speech Recognition for Nigerian Pidgin." In: *ArXiv* abs/2010.11123.

Alleyne, Mervyn (1971). "Acculturation and the cultural matrix of creolization." In: *Pidginization and*, pp. 169–186.

Althobaiti, Maha, Udo Kruschwitz, and Massimo Poesio (Apr. 2014). "Automatic Creation of Arabic Named Entity Annotated Corpus Using Wikipedia." In: *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 106–115. DOI: 10.3115/v1/E14-3012. URL: https://aclanthology.org/E14-3012.

Aralikatte, Rahul, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard (Nov. 2019). "Rewarding Coreference Resolvers for Being Consistent with World Knowledge." 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). Hong Kong, China: Association for Computational Linguistics, pp. 1229–1235. DOI: 10.18653/v1/D19-1118. URL: https://aclanthology.org/D19-1118.

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (Nov. 2020). "Translation Artifacts in Cross-lingual Transfer Learning." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. URL: https://aclanthology.org/2020.emnlp-main.618.

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama (July 2020). "On the Cross-lingual Transferability of Monolingual Representations." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: https://aclanthology.org/2020.acl-main.421.

Athreya, Ram G., Srividya Kona Bansal, Axel-Cyrille Ngonga-Ngomo, and Ricardo Usbeck (2021). "Template-based Question Answering using Recursive Neural Networks." In: *2021 IEEE 15th International Conference on Semantic Computing (ICSC)*, pp. 195–198.

Bajpai, Rajiv, Soujanya Poria, Danyuan Ho, and Erik Cambria (2017). "Developing a concept-level knowledge base for sentiment analysis in Singlish." In: *CoRR* abs/1707.04408. arXiv: 1707.04408. URL: http://arxiv.org/abs/1707.04408.

Baker, Philip and Guillaume Fon Sing (2007). *The making of Mauritian Creole. Analyses diachroniques à partir des textes anciens*. 9. Battlebridge.

Banarescu, Laura, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider (Aug. 2013). "Abstract Meaning Representation for Sembanking." In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Sofia, Bulgaria: Association for Computational Linguistics, pp. 178–186. URL: https://aclanthology.org/W13-2322.

Bartolo, Max, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp (2020). "Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension." In: *Transactions of the Association for Computational Linguistics* 8, pp. 662–678. DOI: 10.1162/tacl_a_00338. URL: https://aclanthology.org/2020.tacl-1.43.

Ben-David, Shai, John Blitzer, Koby Crammer, and Fernando Pereira (2007). "Analysis of Representations for Domain Adaptation." In: *Advances in Neural Information Processing Systems*. Ed. by B. Schölkopf, J. Platt, and T. Hoffman. Vol. 19. MIT Press. URL: https://proceedings.

neurips.cc/paper/2006/file/b1b0432ceafb0ce714426e9114852ac7-Paper.pdf.

Ben-Tal, A., D. D. Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen (2013). "Robust Solutions of Optimization Problems Affected by Uncertain Probabilities." In: *Manag. Sci.* 59, pp. 341–357.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021). "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '21. Virtual Event, Canada: Association for Computing Machinery, 610–623. ISBN: 9781450383097. DOI: 10.1145/3442188.3445922. URL: https://doi.org/10.1145/3442188.3445922.

Berant, Jonathan and Percy Liang (June 2014). "Semantic Parsing via Paraphrasing." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, pp. 1415–1425. DOI: 10.3115/v1/P14-1133. URL: https://aclanthology.org/P14-1133.

Bickerton, Derek (1984). "The Language Bioprogram Hypothesis." In: *Behavioral and brain sciences* 7.2, pp. 173–188.

Bigi, B., B. Caron, and Oyelere S. Abiola (2017). "Developing Resources for Automated Speech Processing of the African Language Naija (Nigerian Pidgin)." In.

Bird, Steven (Dec. 2020). "Decolonising Speech and Language Technology." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313. URL: https://aclanthology.org/2020.coling-main.313.

Björkelund, Anders, Love Hafdell, and Pierre Nugues (June 2009). "Multilingual Semantic Role Labeling." In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Boulder, Colorado: Association for Computational Linguistics, pp. 43–48. URL: https://www.aclweb.org/anthology/W09-1206.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III, and Hanna Wallach (July 2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5454–5476. DOI: 10.18653/v1/2020.acl-main.485. URL: https://aclanthology.org/2020.acl-main.485.

Bohnet, Bernd, Ryan McDonald, Gonçalo Simões, Daniel Andor, Emily Pitler, and Joshua Maynez (July 2018). "Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Aus-

tralia: Association for Computational Linguistics, pp. 2642–2652. DOI: 10.18653/v1/P18-1246. URL: https://aclanthology.org/P18-1246.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai (2016). "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 4356–4364. ISBN: 9781510838819.

Bommasani, Rishi et al. (2021). "On the Opportunities and Risks of Foundation Models." In: *CoRR* abs/2108.07258. arXiv: 2108.07258. URL: https://arxiv.org/abs/2108.07258.

Bornkessel, Ina, Matthias Schlesewsky, Bernard Comrie, and Angela D Friederici (2009). *Semantic role universals and argument linking: Theoretical, typological, and psycholinguistic perspectives*. Vol. 165. Walter de Gruyter.

Budur, Emrah, Rıza Özçelik, Tunga Gungor, and Christopher Potts (Nov. 2020). "Data and Representation for Turkish Natural Language Inference." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 8253–8267. DOI: 10.18653/v1/2020.emnlp-main.662. URL: https://aclanthology.org/2020.emnlp-main.662.

Budzianowski, Pawel, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic (2018). "MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling." In: *CoRR* abs/1810.00278. arXiv: 1810.00278. URL: http://arxiv.org/abs/1810.00278.

Callison-Burch, Chris, Philipp Koehn, Christof Monz, and Omar F. Zaidan, eds. (2011). *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics. URL: http://www.aclweb.org/anthology/W11-21.

Cao, Shulin, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang (May 2022). "KQA Pro: A Dataset with Explicit Compositional Programs for Complex Question Answering over Knowledge Base." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6101–6119. DOI: 10.18653/v1/2022.acl-long.422. URL: https://aclanthology.org/2022.acl-long.422.

Caron, Bernard, Marine Courtin, Kim Gerdes, and Sylvain Kahane (Aug. 2019). "A Surface-Syntactic UD Treebank for Naija." In: *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*. Paris, France: Association for Com-

putational Linguistics, pp. 13–24. DOI: 10.18653/v1/W19-7803. URL: https://www.aclweb.org/anthology/W19-7803.

Carson, Anne (1998). *Autobiography of Red: A Novel in Verse*. New York: Alfred A. Knopf,

Çetinoğlu, Özlem, Sarah Schulz, and Ngoc Thang Vu (Nov. 2016). "Challenges of Computational Processing of Code-Switching." In: *Proceedings of the Second Workshop on Computational Approaches to Code Switching*. Austin, Texas: Association for Computational Linguistics, pp. 1–11. DOI: 10.18653/v1/W16-5801. URL: https://www.aclweb.org/anthology/W16-5801.

Chalkidis, Ilias, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras (May 2022). "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 4310–4330. DOI: 10.18653/v1/2022.acl-long.297. URL: https://aclanthology.org/2022.acl-long.297.

Chau, Ethan C., Lucy H. Lin, and Noah A. Smith (Nov. 2020). "Parsing with Multilingual BERT, a Small Corpus, and a Small Treebank." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1324–1334. DOI: 10.18653/v1/2020.findings-emnlp.118. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.118.

Chen, Chenhua, Alexis Palmer, and Caroline Sporleder (Nov. 2011). "Enhancing Active Learning for Semantic Role Labeling via Compressed Dependency Trees." In: *Proceedings of 5th International Joint Conference on Natural Language Processing*. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, pp. 183–191. URL: https://www.aclweb.org/anthology/I11-1021.

Chen, T. and Kan Min-Yen (2015a). *The National University of Singapore SMS Corpus*.

– (2015b). *The National University of Singapore SMS Corpus*.

Chen, Zhiyu et al. (Nov. 2021). "FinQA: A Dataset of Numerical Reasoning over Financial Data." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 3697–3711. DOI: 10.18653/v1/2021.emnlp-main.300. URL: https://aclanthology.org/2021.emnlp-main.300.

Chiang, David (June 2005). "A Hierarchical Phrase-Based Model for Statistical Machine Translation." In: *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 263–270. DOI: 10.3115/1219840.1219873. URL: https://aclanthology.org/P05-1033.

Choi, DongHyun, Myeong Cheol Shin, EungGyun Kim, and Dong Ryeol Shin (2020). *RYANSQL: Recursively Applying Sketch-based Slot Fillings for Complex Text-to-SQL in Cross-Domain Databases*. arXiv: 2004.03125 [cs.CL].

Chomsky, Noam and Howard Lasnik (2008). "The theory of principles and parameters." In: *1. Halbband*. De Gruyter Mouton, pp. 506–569.

Cirik, Volkan, Eduard H. Hovy, and Louis-Philippe Morency (2016). "Visualizing and Understanding Curriculum Learning for Long Short-Term Memory Networks." In: *CoRR* abs/1611.06204. arXiv: 1611.06204. URL: http://arxiv.org/abs/1611.06204.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Unsupervised Cross-lingual Representation Learning at Scale." In: *CoRR* abs/1911.02116. arXiv: 1911.02116. URL: http://arxiv.org/abs/1911.02116.

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov (July 2020). "Unsupervised Cross-lingual Representation Learning at Scale." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: https://aclanthology.org/2020.acl-main.747.

Conneau, Alexis, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov (2018). "XNLI: Evaluating Cross-lingual Sentence Representations." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485. DOI: 10.18653/v1/D18-1269. URL: https://aclanthology.org/D18-1269.

Corne, Chris (1999). *From French to Creole: The development of new vernaculars in the French colonial world*. Vol. 5. Westminster creolistics.

Cui, Ruixiang, Rahul Aralikatte, Heather C. Lent, and Daniel Hershcovich (2021). "Multilingual Compositional Wikidata Questions." In: *CoRR* abs/2108.03509. arXiv: 2108.03509. URL: https://arxiv.org/abs/2108.03509.

Dabre, Raj, Aneerav Sukhoo, and Pushpak Bhattacharyya (Dec. 2014). "Anou Tradir: Experiences In Building Statistical Machine Translation Systems For Mauritian Languages – Creole, English, French." In: *Proceedings of the 11th International Conference on Natural Language Processing*. Goa, India: NLP Association of India, pp. 82–88. URL: https://aclanthology.org/W14-5113.

Dahl, Deborah A., Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg (1994). "Expanding the Scope of the

ATIS Task: The ATIS-3 Corpus." In: *Proceedings of the Workshop on Human Language Technology*. HLT '94. Plainsboro, NJ: Association for Computational Linguistics, 43–48. ISBN: 1558603573. DOI: 10.3115/1075812.1075823. URL: https://doi.org/10.3115/1075812.1075823.

Damonte, Marco and Emilio Monti (2021). "One Semantic Parser to Parse Them All: Sequence to Sequence Multi-Task Learning on Semantic Parsing Datasets." In: *CoRR* abs/2106.04476. arXiv: 2106.04476. URL: https://arxiv.org/abs/2106.04476.

Daval-Markussen, Aymeric and Peter Bakker (2012). "Explorations in creole research with phylogenetic tools." In: *EACL 2012*.

DeCamp, David (1971). *Toward a generative analysis of a post-creole speech continuum*. Cambridge University Press.

DeGraff, Michael (2005a). "o Creole languages constitute an exceptional typological class?" In: *Revue française de linguistique appliquée* 10.1, pp. 11–24.

DeGraff, Michel (2001). "On the origin of creoles: A Cartesian critique of neo-Darwinian linguistics." In: *Linguistic Typology* 5.2/3, pp. 213–310.

– (2003). "Against creole exceptionalism." In: *Language* 79.2, pp. 391–410.

– (2005b). "Linguists' most dangerous myth: The fallacy of Creole Exceptionalism." In: *Language in society* 34.4, pp. 533–591.

– (2005c). "Linguists' most dangerous myth: The fallacy of Creole Exceptionalism." In: *Language in Society* 34, pp. 533 –591.

– (2007). "Kreyòl Ayisyen, or Haitian Creole (Creole French)." In: *Comparative creole syntax: Parallel outlines of* 18, pp. 101–126.

– (2020). *Toward racial justice in linguistics: The case of Creole studies (Response to Charity Hudley et al.)*. DOI: doi:10.1353/lan.2020.0080.

Deshpande, Ameet, Partha Talukdar, and Karthik Narasimhan (2021). *When is BERT Multilingual? Isolating Crucial Ingredients for Cross-lingual Transfer*. arXiv: 2110.14782 [cs.CL].

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (June 2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://www.aclweb.org/anthology/N19-1423.

Diefenbach, Dennis, Thomas Pellissier Tanon, Kamal Deep Singh, and Pierre Maret (2017). "Question Answering Benchmarks for Wikidata." In: *SEMWEB*.

Doan, Long, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen (Nov. 2021). "PhoMT: A High-Quality and Large-Scale Benchmark Dataset for Vietnamese-English Machine

Translation." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4495–4503. DOI: 10.18653/v1/2021.emnlp-main.369. URL: https://aclanthology.org/2021.emnlp-main.369.

Doğruöz, A. Seza, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio (Aug. 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)*. Online: Association for Computational Linguistics, pp. 1654–1666. DOI: 10.18653/v1/2021.acl-long.131. URL: https://aclanthology.org/2021.acl-long.131.

Dozat, Timothy, Peng Qi, and Christopher D. Manning (Aug. 2017). "Stanford's Graph-based Neural Dependency Parser at the CoNLL 2017 Shared Task." In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, pp. 20–30. DOI: 10.18653/v1/K17-3002. URL: https://www.aclweb.org/anthology/K17-3002.

Dryer, Matthew S. and Martin Haspelmath, eds. (2013). *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://wals.info/.

Dufter, Philipp, Martin Schmitt, and Hinrich Schütze (Dec. 2020). "Increasing Learning Efficiency of Self-Attention Networks through Direct Position Interactions, Learnable Temperature, and Convoluted Attention." In: *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3630–3636. DOI: 10.18653/v1/2020.coling-main.324. URL: https://aclanthology.org/2020.coling-main.324.

Dufter, Philipp and Hinrich Schütze (Nov. 2020). "Identifying Elements Essential for BERT's Multilinguality." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4423–4437. DOI: 10.18653/v1/2020.emnlp-main.358. URL: https://aclanthology.org/2020.emnlp-main.358.

Elazar, Yanai, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg (2021). "Measuring and Improving Consistency in Pretrained Language Models." In: *Transactions of the Association for Computational Linguistics* 9, pp. 1012–1031. DOI: 10.1162/tacl_a_00410. URL: https://aclanthology.org/2021.tacl-1.60.

Eric, Mihail and Christopher D. Manning (2017). "Key-Value Retrieval Networks for Task-Oriented Dialogue." In: *CoRR* abs/1705.05414. arXiv: 1705.05414. URL: http://arxiv.org/abs/1705.05414.

Fatima, Mehwish and Michael Strube (Nov. 2021). "A Novel Wikipedia based Dataset for Monolingual and Cross-Lingual Summarization." In: *Proceedings of the Third Workshop on New Frontiers in Summarization*. Online and in Dominican Republic: Association for Computational Linguistics, pp. 39–50. DOI: 10.18653/v1/2021.newsum-1.5. URL: https://aclanthology.org/2021.newsum-1.5.

Fei, Geli and Bing Liu (June 2016). "Breaking the Closed World Assumption in Text Classification." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, pp. 506–514. DOI: 10.18653/v1/N16-1061. URL: https://www.aclweb.org/anthology/N16-1061.

Feldman, Vitaly and Chiyuan Zhang (2020). "What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation." In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 2881–2891. URL: https://proceedings.neurips.cc/paper/2020/file/1e14bfe2714193e7af5abc64ecbd6b46-Paper.pdf.

Finegan-Dollak, Catherine, Jonathan K. Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev (July 2018). "Improving Text-to-SQL Evaluation Methodology." In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 351–360. DOI: 10.18653/v1/P18-1033. URL: https://aclanthology.org/P18-1033.

Furman, Gregory and G. Nitschke (2020). "Evolving an artificial creole." In: *Proceedings of the 2020 Genetic and Evolutionary Computation Conference Companion*.

Goel, Karan, Nazneen Fatema Rajani, Jesse Vig, Samson Tan, Jason Wu, Stephan Zheng, Caiming Xiong, Mohit Bansal, and Christopher Ré (2021). "Robustness Gym: Unifying the NLP Evaluation Landscape." In: *CoRR* abs/2101.04840. arXiv: 2101.04840. URL: https://arxiv.org/abs/2101.04840.

Goot, Rob van der, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank (Apr. 2021a). "Massive Choice, Ample Tasks (MaChAmp): A Toolkit for Multi-task Learning in NLP." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 176–197. URL: https://www.aclweb.org/anthology/2021.eacl-demos.22.

Goot, Rob van der, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank (2021b). *Massive Choice, Ample Tasks (MaChAmp):A Toolkit for Multi-task Learning in NLP*.

Graham, Yvette, Barry Haddow, and Philipp Koehn (Nov. 2020). "Statistical Power and Translationese in Machine Translation Evaluation." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 72–81. DOI: 10.18653/v1/2020.emnlp-main.6. URL: https://aclanthology.org/2020.emnlp-main.6.

Gruber, Jeffrey Steven (1965). "Studies in lexical relations." PhD thesis. Massachusetts Institute of Technology.

Guzmán, Francisco, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato (Nov. 2019). "The FLORES Evaluation Datasets for Low-Resource Machine Translation: Nepali–English and Sinhala–English." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 6098–6111. DOI: 10.18653/v1/D19-1632. URL: https://aclanthology.org/D19-1632.

Hagemeijer, Tjerk, Michel Généreux, Iris Hendrickx, Amália Mendes, Abigail Tiny, and Armando Zamora (May 2014). "The Gulf of Guinea Creole Corpora." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 523–529. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/438_Paper.pdf.

Hashimoto, Tatsunori, Megha Srivastava, Hongseok Namkoong, and Percy Liang (2018). "Fairness Without Demographics in Repeated Loss Minimization." In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1929–1938. URL: http://proceedings.mlr.press/v80/hashimoto18a.html.

He, Luheng, Kenton Lee, Mike Lewis, and Luke Zettlemoyer (July 2017). "Deep Semantic Role Labeling: What Works and What's Next." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 473–483. DOI: 10.18653/v1/P17-1044. URL: https://www.aclweb.org/anthology/P17-1044.

Henri, Fabiola, Gregory Stump, and Delphine Tribout (2020). "Derivation and the morphological complexity of three French-based creoles." In: *The Complexities of Morphology*. Oxford University Press, pp. 105–135.

Hershcovich, Daniel, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend (June 2019). "SemEval-2019 Task 1: Cross-lingual Semantic Parsing with UCCA." In: *Proceedings of the 13th International Workshop on Semantic Evaluation*. Minneapolis, Minnesota, USA: Association for Computational Linguistics, pp. 1–10. DOI: 10.18653/v1/S19-2001. URL: https://www.aclweb.org/anthology/S19-2001.

Hershcovich, Daniel et al. (May 2022). "Challenges and Strategies in Cross-Cultural NLP." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 6997–7013. DOI: 10.18653/v1/2022.acl-long.482. URL: https://aclanthology.org/2022.acl-long.482.

Hewavitharana, Sanjika, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel (July 2011). "CMU Haitian Creole-English Translation System for WMT 2011." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 386–392. URL: https://www.aclweb.org/anthology/W11-2146.

Hinrichs, Lars (2006). *Codeswitching on the Web*. John Benjamins Amsterdam.

Ho, Danyuan, Diyana Hamzah, Soujanya Poria, and Erik Cambria (2018a). "Singlish SenticNet: A Concept-Based Sentiment Resource for Singapore English." In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1285–1291. DOI: 10.1109/SSCI.2018.8628796.

– (2018b). "Singlish SenticNet: A Concept-Based Sentiment Resource for Singapore English." In: *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1285–1291. DOI: 10.1109/SSCI.2018.8628796.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Holm, John (1982). *Central American English*. Vol. 2. John Benjamins Publishing.

– (2000). "Social factors." In: *An Introduction to Pidgins and Creoles*. Cambridge Textbooks in Linguistics. Cambridge University Press, 68–105. DOI: 10.1017/CBO9781139164153.006.

Hovy, Eduard H., Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel (2006a). "OntoNotes: The 90% Solution." In: *NAACL*.

Hovy, Eduard, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel (June 2006b). "OntoNotes: The 90% Solution." In: *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*. New York City, USA: Association for Computational Linguistics, pp. 57–60. URL: https://www.aclweb.org/anthology/N06-2015.

Hu, Chang, P. Resnik, Y. Kronrod, Vladimir Eidelman, Olivia Buzek, and B. Bederson (2011a). "The Value of Monolingual Crowdsourcing in a Real-World Translation Scenario: Simulation using Haitian Creole Emergency SMS Messages." In: *WMT@EMNLP*.

Hu, Chang, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson (2011b). "The Value of Monolingual Crowdsourcing in a Real-World Translation Scenario: Simulation using Haitian Creole Emergency SMS Messages." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 399–404. URL: http://www.aclweb.org/anthology/W11-2148.

– (July 2011c). "The Value of Monolingual Crowdsourcing in a Real-World Translation Scenario: Simulation using Haitian Creole Emergency SMS Messages." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 399–404. URL: https://aclanthology.org/W11-2148.

Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson (2020). "XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization." In: *CoRR* abs/2003.11080. arXiv: 2003.11080. URL: https://arxiv.org/abs/2003.11080.

Huson, Daniel H and David Bryant (2006). "Application of phylogenetic networks in evolutionary studies." In: *Molecular biology and evolution* 23.2, pp. 254–267.

Iyer, Srinivasan, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer (July 2017). "Learning a Neural Semantic Parser from User Feedback." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 963–973. DOI: 10.18653/v1/P17-1089. URL: https://aclanthology.org/P17-1089.

Iyyer, Mohit, John Wieting, Kevin Gimpel, and Luke Zettlemoyer (June 2018). "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1875–1885. DOI: 10.18653/v1/N18-1170. URL: https://aclanthology.org/N18-1170.

Jansson, Fredrik, Mikael Parkvall, and Pontus Strimling (2015). "Modeling the Evolution of Creoles." In: *Language Dynamics and Change* 5.1, pp. 1 –51. DOI: https://doi.org/10.1163/22105832-00501005. URL: https://brill.com/view/journals/ldc/5/1/article-p1_1.xml.

Jia, Robin and Percy Liang (Aug. 2016). "Data Recombination for Neural Semantic Parsing." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 12–22. DOI: 10.18653/v1/P16-1002. URL: https://aclanthology.org/P16-1002.

Jiao, Xiaoqi, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu (Nov. 2020). "TinyBERT: Distilling BERT for Natural Language Understanding." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4163–4174. DOI: 10.18653/v1/2020.findings-emnlp.372. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.372.

Joshi, Pratik M., Sebastin Santy, Amarjit Budhiraja, Kalika Bali, and Monojit Choudhury (2020a). "The State and Fate of Linguistic Diversity and Inclusion in the NLP World." In: *ACL*.

Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury (July 2020b). "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*. Online: Association for Computational Linguistics, pp. 6282–6293. DOI: 10.18653/v1/2020.acl-main.560. URL: https://www.aclweb.org/anthology/2020.acl-main.560.

K, Karthikeyan, Zihan Wang, Stephen Mayhew, and Dan Roth (2020). "Cross-Lingual Ability of Multilingual BERT: An Empirical Study." In: *ArXiv* abs/1912.07840.

Kannan Ravi, Manoj Prabhakar, Kuldeep Singh, Isaiah Onando Mulang', Saeedeh Shekarpour, Johannes Hoffart, and Jens Lehmann (Apr. 2021). "CHOLAN: A Modular Approach for Neural Entity Linking on Wikipedia and Wikidata." In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 504–514. DOI: 10.18653/v1/2021.eacl-main.40. URL: https://aclanthology.org/2021.eacl-main.40.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). "Scaling Laws for Neural Language Models." In: *CoRR* abs/2001.08361. arXiv: 2001.08361. URL: https://arxiv.org/abs/2001.08361.

Keung, Phillip, Yichao Lu, Julian Salazar, and Vikas Bhardwaj (Nov. 2020). "Don't Use English Dev: On the Zero-Shot Cross-Lingual Evaluation of Contextual Embeddings." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 549–554. DOI: 10.18653/v1/2020.emnlp-main.40. URL: https://aclanthology.org/2020.emnlp-main.40.

Kiela, Douwe et al. (2021). "Dynabench: Rethinking Benchmarking in NLP." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou. Association for Computational Linguistics, pp. 4110–4124. DOI: 10.18653/v1/2021.naacl-main.324. URL: https://doi.org/10.18653/v1/2021.naacl-main.324.

Koh, Pang Wei et al. (2020). "WILDS: A Benchmark of in-the-Wild Distribution Shifts." In: *CoRR* abs/2012.07421. arXiv: 2012.07421. URL: https://arxiv.org/abs/2012.07421.

Koh, Pang Wei et al. (2021). *WILDS: A Benchmark of in-the-Wild Distribution Shifts*. arXiv: 2012.07421 [cs.LG].

Korablinov, Vladislav and Pavel Braslavski (2020). "RuBQ: A Russian Dataset for Question Answering over Wikidata." In: *CoRR* abs/2005.10659. arXiv: 2005.10659. URL: https://arxiv.org/abs/2005.10659.

Kouwenberg, Silvia and John Victor Singler (2009). *The handbook of pidgin and creole studies*. John Wiley & Sons.

Kriegel, Sibylle (2016). "2014. Pidgins and Creoles beyond Africa-Europe Encounters, written by Isabelle Buchstaller, Anders Holmberg and Mohammed Almoaily." In: *Journal of Language Contact* 9.3, pp. 576 –579. DOI: https://doi.org/10.1163/19552629-00903007. URL: https://brill.com/view/journals/jlc/9/3/article-p576_7.xml.

Kwiatkowski, Tom, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer (Oct. 2013). "Scaling Semantic Parsers with On-the-Fly Ontology Matching." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 1545–1556. URL: https://aclanthology.org/D13-1161.

Kwiatkowski, Tom et al. (2019). "Natural Questions: A Benchmark for Question Answering Research." In: *Transactions of the Association for Computational Linguistics* 7, pp. 452–466. DOI: 10.1162/tacl_a_00276. URL: https://aclanthology.org/Q19-1026.

Lake, Brenden M. and Marco Baroni (2017). "Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks." In: *CoRR* abs/1711.00350. arXiv: 1711.00350. URL: http://arxiv.org/abs/1711.00350.

Lauscher, Anne, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš (Nov. 2020). "From Zero to Hero: On the Limitations of Zero-Shot Language Transfer with Multilingual Transformers." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguis-

tics, pp. 4483–4499. DOI: 10.18653/v1/2020.emnlp-main.363. URL: https://aclanthology.org/2020.emnlp-main.363.

Lent, Heather, Emanuele Bugliarello, Miryam de Lhoneux, Chen Qiu, and Anders Søgaard (Nov. 2021a). "On Language Models for Creoles." In: *Proceedings of the 25th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 58–71. DOI: 10.18653/v1/2021.conll-1.5. URL: https://aclanthology.org/2021.conll-1.5.

Lent, Heather, Emanuele Bugliarello, and Anders Søgaard (May 2022). "Ancestor-to-Creole Transfer is Not a Walk in the Park." In: *Proceedings of the Third Workshop on Insights from Negative Results in NLP*. Dublin, Ireland: Association for Computational Linguistics, pp. 68–74. DOI: 10.18653/v1/2022.insights-1.9. URL: https://aclanthology.org/2022.insights-1.9.

Lent, Heather, Kelechi Ogueji, Miryam de Lhoneux, Orevaoghene Ahia, and Anders Søgaard (2022). "What a Creole Wants, What a Creole Needs." In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 6439–6449. URL: https://aclanthology.org/2022.lrec-1.691.

Lent, Heather, Semih Yavuz, Tao Yu, Tong Niu, Yingbo Zhou, Dragomir Radev, and Xi Victoria Lin (Nov. 2021b). "Testing Cross-Database Semantic Parsers With Canonical Utterances." In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 73–83. DOI: 10.18653/v1/2021.eval4nlp-1.8. URL: https://aclanthology.org/2021.eval4nlp-1.8.

Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer (July 2020a). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

Lewis, Patrick, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk (July 2020b). "MLQA: Evaluating Cross-lingual Extractive Question Answering." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7315–7330. DOI: 10.18653/v1/2020.acl-main.653. URL: https://aclanthology.org/2020.acl-main.653.

Lin, Xi Victoria, Richard Socher, and Caiming Xiong (Nov. 2020). "Bridging Textual and Tabular Data for Cross-Domain Text-to-SQL Semantic Parsing." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Lin-

guistics, pp. 4870–4888. DOI: 10.18653/v1/2020.findings-emnlp.438. URL: https://aclanthology.org/2020.findings-emnlp.438.

Lin, Ying, Xiaoman Pan, Aliya Deri, Heng Ji, and Kevin Knight (Aug. 2016). "Leveraging Entity Linking and Related Language Projection to Improve Name Transliteration." In: *Proceedings of the Sixth Named Entity Workshop*. Berlin, Germany: Association for Computational Linguistics, pp. 1–10. DOI: 10.18653/v1/W16-2701. URL: https://aclanthology.org/W16-2701.

Littell, Patrick, Kartik Goyal, David R. Mortensen, Alexa Little, Chris Dyer, and Lori Levin (Dec. 2016). "Named Entity Recognition for Linguistic Rapid Response in Low-Resource Languages: Sorani Kurdish and Tajik." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 998–1006. URL: https://aclanthology.org/C16-1095.

Liu, Dayiheng, Yeyun Gong, Jie Fu, Yu Yan, Jiusheng Chen, Daxin Jiang, Jiancheng Lv, and Nan Duan (July 2020). "RikiNet: Reading Wikipedia Pages for Natural Question Answering." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6762–6771. DOI: 10.18653/v1/2020.acl-main.604. URL: https://aclanthology.org/2020.acl-main.604.

Liu, Fangyu, Emanuele Bugliarello, E. Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott (2021a). "Visually Grounded Reasoning across Languages and Cultures." In: *ArXiv* abs/2109.13238.

Liu, Pengfei, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaichen Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, and Graham Neubig (2021b). "EXPLAINABOARD: An Explainable Leaderboard for NLP." In: *CoRR* abs/2104.06387. arXiv: 2104.06387. URL: https://arxiv.org/abs/2104.06387.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].

Lo, Siaw Ling, Erik Cambria, Raymond Chiong, and David Cornforth (2016). "A multilingual semi-supervised approach in deriving Singlish sentic patterns for polarity detection." In: *Knowledge-Based Systems* 105, pp. 236–247. ISSN: 0950-7051. DOI: https://doi.org/10.1016/j.knosys.2016.04.024. URL: https://www.sciencedirect.com/science/article/pii/S0950705116300764.

Marcheggiani, Diego and Ivan Titov (2017). "Encoding Sentences with Graph Convolutional Networks for Semantic Role Labeling." In: *EMNLP*.

Mayer, Thomas and Michael Cysouw (May 2014). "Creating a massively parallel Bible corpus." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reyk-

javik, Iceland: European Language Resources Association (ELRA), pp. 3158–3163. URL: http://www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf.

McWhorter, John H. (1998). "Identifying the Creole Prototype: Vindicating a Typological Class." In: *Language* 74.4, pp. 788–818. ISSN: 00978507, 15350665. URL: http://www.jstor.org/stable/417003.

Michaelis, Susanne Maria, Philippe Maurer, Martin Haspelmath, and Magnus Huber, eds. (2013). *APiCS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. URL: https://apics-online.info/.

Mielke, Sabrina J., Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner (2019). "What Kind of Language Is Hard to Language-Model?" In: *ACL*.

Migge, Bettina (2020). "(A Review of) The Creole Debate, written by John H. McWhorter." In: *Journal of Language Contact* 12.3, pp. 857–863. DOI: https://doi.org/10.1163/19552629-01203009. URL: https://brill.com/view/journals/jlc/12/3/article-p857_857.xml.

Millour, Alice and Karën Fort (May 2020). "Text Corpora and the Challenge of Newly Written Languages." 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)*. Marseille, France: European Language Resources association, pp. 111–120. ISBN: 979-10-95546-35-1. URL: https://aclanthology.org/2020.sltu-1.15.

Mirzakhalov, Jamshidbek et al. (Nov. 2021). "A Large-Scale Study of Machine Translation in Turkic Languages." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 5876–5890. DOI: 10.18653/v1/2021.emnlp-main.475. URL: https://aclanthology.org/2021.emnlp-main.475.

Mufwene, Salikoko (2014). "Short notes: The case was never closed. McWhorter misinterprets the ecological approach to the emergence of creole." In: *Journal of Pidgin and Creole Languages* 29.1, pp. 157–171.

Muhammad, Shamsuddeen Hassan et al. (2022). *NaijaSenti: A Nigerian Twitter Sentiment Corpus for Multilingual Sentiment Analysis*. DOI: 10.48550/ARXIV.2201.08277. URL: https://arxiv.org/abs/2201.08277.

Muller, Benjamin, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah (June 2021). "When Being Unseen from mBERT is just the Beginning: Handling New Languages With Multilingual Language Models." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Lin-

guistics, pp. 448–462. DOI: 10.18653/v1/2021.naacl-main.38. URL: https://aclanthology.org/2021.naacl-main.38.

Munro, Robert (2010). "Crowdsourced translation for emergency response in haiti: the global collaboration of local knowledge." In: *In Relief 2.0 in Haiti*.

– (2013). "Crowdsourcing and the crisis-affected community - Lessons learned and looking forward from Mission 4636." In: *Inf. Retr.* 16.2, pp. 210–266. URL: http://dblp.uni-trier.de/db/journals/ir/ir16.html#Munro13.

Murawaki, Yugo (2016). "Statistical Modeling of Creole Genesis." In: *NAACL*.

Muysken, Pieter and Norval Smith (1986). *Substrata versus universals in creole genesis: papers from the Amsterdam Creole Workshop, April 1985*. Vol. 1. John Benjamins Publishing.

Nakamura, Makoto, Takashi Hashimoto, and Satoshi Tojo (2009). "Prediction of Creole Emergence in Spatial Language Dynamics." In: *Language and Automata Theory and Applications*. Ed. by Adrian Horia Dediu, Armand Mihai Ionescu, and Carlos Martín-Vide. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 614–625. ISBN: 978-3-642-00982-2.

Ndubuisi-Obi, Innocent, S. Ghosh, and David Jurgens (2019a). "Wetin dey with these comments? Modeling Sociolinguistic Factors Affecting Code-switching Behavior in Nigerian Online Discussions." In: *ACL*.

Ndubuisi-Obi, Innocent, Sayan Ghosh, and David Jurgens (2019b). "Wétin dey with these comments? Modeling Sociolinguistic Factors Affecting Code-switching Behavior, in Nigerian Online Discussions." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Nekoto, Wilhelmina et al. (Nov. 2020). "Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 2144–2160. DOI: 10.18653/v1/2020.findings-emnlp.195. URL: https://aclanthology.org/2020.findings-emnlp.195.

Nie, Binling, Ruixue Ding, Pengjun Xie, Fei Huang, Chen Qian, and Luo Si (2021). "Knowledge-aware Named Entity Recognition with Alleviating Heterogeneity." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.15, pp. 13595–13603. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17603.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman (May 2020a). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection." English. In: *LREC*. Marseille, France: European Language Resources Associa-

tion, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://www.aclweb.org/anthology/2020.lrec-1.497.

– (May 2020b). "Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection." English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4034–4043. ISBN: 979-10-95546-34-4. URL: https://aclanthology.org/2020.lrec-1.497.

Nothman, Joel, James R. Curran, and Tara Murphy (Dec. 2008). "Transforming Wikipedia into Named Entity Training Data." In: *Proceedings of the Australasian Language Technology Association Workshop 2008*. Hobart, Australia, pp. 124–132. URL: https://aclanthology.org/U08-1016.

Ogueji, Kelechi and Orevaoghene Ahia (2019). "PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English." In: *ArXiv* abs/1912.03444.

Ogueji, Kelechi, Yuxin Zhu, and Jimmy Lin (Nov. 2021). "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages." In: *Proceedings of the 1st Workshop on Multilingual Representation Learning*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 116–126. DOI: 10.18653/v1/2021.mrl-1.11. URL: https://aclanthology.org/2021.mrl-1.11.

Oren, Yonatan, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang (Nov. 2019a). "Distributionally Robust Language Modeling." 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4227–4237. DOI: 10.18653/v1/D19-1432. URL: https://www.aclweb.org/anthology/D19-1432.

Oren, Yonatan, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang (2019b). "Distributionally robust language modeling." In: *arXiv preprint arXiv:1909.02060*.

Östling, Robert and Jörg Tiedemann (Apr. 2017). "Continuous multilinguality with language vectors." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics, pp. 644–649. URL: https://aclanthology.org/E17-2102.

Oyewusi, Wuraola Fisayo, Olubayo Adekanmbi, and Olalekan Akinsande (2020). "Semantic Enrichment of Nigerian Pidgin English for Contextual Sentiment Classification." In: *ArXiv* abs/2003.12450.

Oyewusi, Wuraola Fisayo, Olubayo Adekanmbi, Ife Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa (2021a). "NaijaNER : Comprehensive Named Entity Recognition for 5 Nigerian Languages." In: *ArXiv* abs/2105.00810.

Oyewusi, Wuraola Fisayo, Olubayo Adekanmbi, Ifeoma Okoh, Vitus Onuigwe, Mary Idera Salami, Opeyemi Osakuade, Sharon Ibejih, and Usman Abdullahi Musa (2021b). *NaijaNER : Comprehensive Named Entity Recognition for 5 Nigerian Languages*. arXiv: 2105.00810 [cs.CL].

Padó, Sebastian, Marco Pennacchiotti, and Caroline Sporleder (Aug. 2008). "Semantic Role Assignment for Event Nominalisations by Leveraging Verbal Data." In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Manchester, UK: Coling 2008 Organizing Committee, pp. 665–672. URL: https://www.aclweb.org/anthology/C08-1084.

Pal, Riya and Dipti Sharma (Aug. 2019). "A Dataset for Semantic Role Labelling of Hindi-English Code-Mixed Tweets." In: *Proceedings of the 13th Linguistic Annotation Workshop*. Florence, Italy: Association for Computational Linguistics, pp. 178–188. DOI: 10.18653/v1/W19-4020. URL: https://www.aclweb.org/anthology/W19-4020.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury (2005). "The Proposition Bank: An Annotated Corpus of Semantic Roles." In: *Computational Linguistics* 31.1, pp. 71–106. DOI: 10.1162/0891201053630264. URL: https://www.aclweb.org/anthology/J05-1004.

Pan, Xiaoman, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji (July 2017). "Cross-lingual Name Tagging and Linking for 282 Languages." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1946–1958. DOI: 10.18653/v1/P17-1178. URL: https://aclanthology.org/P17-1178.

Parkvall, Mikael (2008). "The simplicity of creoles in a cross-linguistic perspective." In.

Parkvall, Mikael et al. (2008). "The simplicity of creoles in a cross-linguistic perspective." In: *Language complexity: Typology, contact, change*, pp. 265–285.

Paszke, Adam et al. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library." In: *CoRR* abs/1912.01703. arXiv: 1912.01703. URL: http://arxiv.org/abs/1912.01703.

Patrick, Peter (1999). *Urban Jamaican Creole. Variation in the Mesolect.* Varieties of English Around the World G17. Amsterdam & Phildadelphia: John Benjamins Publishing Co. URL: http://www.benjamins.com/cgi-bin/t_bookview.cgi?bookid=VEAWG17.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (Oct. 2014). "GloVe: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: https://www.aclweb.org/anthology/D14-1162.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (June 2018). "Deep

Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

Pfeiffer, Jonas, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder (Nov. 2021). "UNKs Everywhere: Adapting Multilingual Language Models to New Scripts." In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 10186–10203. URL: https://aclanthology.org/2021.emnlp-main.800.

Plag, I. (2009). "Creoles as interlanguages : Phonology." In: *Journal of Pidgin and Creole Languages* 24, pp. 119–138.

Pratapa, Adithya, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali (July 2018). "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)*. Melbourne, Australia: Association for Computational Linguistics, pp. 1543–1553. DOI: 10.18653/v1/P18-1143. URL: https://www.aclweb.org/anthology/P18-1143.

Purschke, Christoph (2021). "Crowdscapes. Participatory research and the collaborative (re) construction of linguistic landscapes with Lingscape." In: *Linguistics Vanguard* 7.s1.

R. Costa-jussà, Marta and Rafael E. Banchs (2011). "The BM-I2R Haitian-Créole-to-English translation system description for the WMT 2011 evaluation campaign." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 452–456. URL: http://www.aclweb.org/anthology/W11-2156.

Rachman, Valdi, Rahmad Mahendra, Alfan Farizki Wicaksono, Ahmad Rizqi Meydiarso, and Fariz Ikhwantri (2018). "Semantic Role Labeling in Conversational Chat using Deep Bi-Directional Long Short-Term Memory Networks with Attention Mechanism." In: *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*. Hong Kong: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/Y18-1064.

Radhakrishnan, Karthik, Arvind Srikantan, and Xi Victoria Lin (2020). "ColloQL: Robust Cross-Domain Text-to-SQL Over Search Queries." In: *CoRR* abs/2010.09927. arXiv: 2010.09927. URL: https://arxiv.org/abs/2010.09927.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016a). "SQuAD: 100, 000+ Questions for Machine Comprehension

of Text." In: *CoRR* abs/1606.05250. arXiv: 1606.05250. URL: http://arxiv.org/abs/1606.05250.

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (Nov. 2016b). "SQuAD: 100,000+ Questions for Machine Comprehension of Text." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: https://aclanthology.org/D16-1264.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh (July 2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4902–4912. DOI: 10.18653/v1/2020.acl-main.442. URL: https://aclanthology.org/2020.acl-main.442.

Richardson, Matthew, Christopher J.C. Burges, and Erin Renshaw (Oct. 2013). "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text." In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA: Association for Computational Linguistics, pp. 193–203. URL: https://aclanthology.org/D13-1020.

Rickford, John R (1987). *Dimensions of a Creole continuum: History, texts & linguistic analysis of Guyanese Creole*. Stanford University Press.

Ringgaard, Michael, Rahul Gupta, and Fernando C Pereira (2017). "SLING: A framework for frame semantic parsing." In: *ArXiv* abs/1710.07032.

Sagawa, Shiori, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang (2019). "Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization." In: *CoRR* abs/1911.08731. arXiv: 1911.08731. URL: http://arxiv.org/abs/1911.08731.

Saha, Amrita, Vardaan Pahuja, Mitesh M. Khapra, Karthik Sankaranarayanan, and Sarath Chandar (2018). "Complex Sequential Question Answering: Towards Learning to Converse over Linked Question Answer Pairs with a Knowledge Graph." In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. AAAI'18/IAAI'18/EAAI'18. New Orleans, Louisiana, USA: AAAI Press. ISBN: 978-1-57735-800-8.

Salazar, Julian, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff (July 2020). "Masked Language Model Scoring." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2699–2712. DOI: 10.18653/v1/2020.acl-main.240. URL: https://www.aclweb.org/anthology/2020.acl-main.240.

Schang, Emmanuel, Jean-Louis Rougé, Iris Eshkol, and Mélanie Petit (2005). "CreolData: A Lexical Database on Creole Languages." In: *Revue francaise de linguistique appliquee* 10.1, pp. 65–76.

Scheirer, Walter J., Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult (2013). "Toward Open Set Recognition." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35.7, pp. 1757–1772. DOI: 10.1109/TPAMI.2012.256.

Sebba, Mark (1997). *Contact languages: Pidgins and creoles*. Macmillan International Higher Education.

– (1998). "Phonology meets ideology: the meaning of orthographic practices in British Creole." In: *Language problems and language planning* 22.1, pp. 19–47.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (Aug. 2016). "Neural Machine Translation of Rare Words with Subword Units." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: https://aclanthology.org/P16-1162.

Sessarego, Sandro (2020). "Not all grammatical features are robustly transmitted during the emergence of creoles." In: *Humanities and Social Sciences Communications* 7, pp. 1–8.

Shah-Sanghavi, Payal Kushal (2017). "Should creoles be made official languages and / or media of instruction in countries where they are the first language of the majority of the population?" In: *IOSR Journal Of Humanities And Social Science* 22, pp. 19–25.

Shaw, Peter, Ming-Wei Chang, Panupong Pasupat, and Kristina Toutanova (2021). "Compositional Generalization and Natural Language Variation: Can a Semantic Parsing Approach Handle Both?" In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Association for Computational Linguistics, pp. 922–938. DOI: 10.18653/v1/2021.acl-long.75. URL: https://doi.org/10.18653/v1/2021.acl-long.75.

Shi, Peng and Jimmy Lin (2019). *Simple BERT Models for Relation Extraction and Semantic Role Labeling*. arXiv: 1904.05255 [cs.CL].

Shi, Peng, Patrick Ng, Zhiguo Wang, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Cicero Nogueira dos Santos, and Bing Xiang (2020). *Learning Contextual Representations for Semantic Parsing with Generation-Augmented Pre-Training*. arXiv: 2012.10309 [cs.CL].

Shin, Joonbo, Yoonhyung Lee, and Kyomin Jung (2019). "Effective Sentence Scoring Method Using BERT for Speech Recognition." In: *Proceedings of The Eleventh Asian Conference on Machine Learning*. Ed. by Wee Sun Lee and Taiji Suzuki. Vol. 101. Proceedings

of Machine Learning Research. Nagoya, Japan: PMLR, pp. 1081–1093. URL: http://proceedings.mlr.press/v101/shin19a.html.

Siegel, Jeff (1999). "Stigmatized and standardized varieties in the classroom: Interference or separation?" In: *Tesol Quarterly* 33.4, pp. 701–728.

Singh, Jasdeep, Bryan McCann, Richard Socher, and Caiming Xiong (Nov. 2019). "BERT is Not an Interlingua and the Bias of Tokenization." In: *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*. Hong Kong, China: Association for Computational Linguistics, pp. 47–55. DOI: 10.18653/v1/D19-6106. URL: https://aclanthology.org/D19-6106.

Slone, Thomas H (2001). *One Thousand One Papua New Guinean Nights: Tales form 1972-1985*. Vol. 1. Masalai Press.

Stanovsky, Gabriel, Julian Michael, Luke Zettlemoyer, and Ido Dagan (June 2018). "Supervised Open Information Extraction." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 885–895. DOI: 10.18653/v1/N18-1081. URL: https://www.aclweb.org/anthology/N18-1081.

Sternberg, Robert and Jacqueline Leighton (2004). *The Nature of Reasoning*. Cambridge University Press.

Strubell, Emma, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum (2018). "Linguistically-Informed Self-Attention for Semantic Role Labeling." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 5027–5038. DOI: 10.18653/v1/D18-1548. URL: https://www.aclweb.org/anthology/D18-1548.

Stymne, Sara (2011). "Spell Checking Techniques for Replacement of Unknown Words and Data Cleaning for Haitian Creole SMS Translation." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Edinburgh, Scotland: Association for Computational Linguistics, pp. 470–477. URL: http://www.aclweb.org/anthology/W11-2159.

Suhr, Alane, Ming-Wei Chang, Peter Shaw, and Kenton Lee (July 2020). "Exploring Unexplored Generalization Challenges for Cross-Database Semantic Parsing." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8372–8388. DOI: 10.18653/v1/2020.acl-main.742. URL: https://aclanthology.org/2020.acl-main.742.

Susanto, Raymond Hendy, Ohnmar Htun, and Liling Tan (Nov. 2019). "Sarah's Participation in WAT 2019." In: *Proceedings of the 6th Workshop on Asian Translation*. Hong Kong, China: Association for Com-

putational Linguistics, pp. 152–158. DOI: 10.18653/v1/D19-5219. URL: https://aclanthology.org/D19-5219.

Tan, Samson, Shafiq Joty, Lav Varshney, and Min-Yen Kan (Nov. 2020). "Mind Your Inflections! Improving NLP for Non-Standard Englishes with Base-Inflection Encoding." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 5647–5663. DOI: 10.18653/v1/2020.emnlp-main.455. URL: https://aclanthology.org/2020.emnlp-main.455.

Thomason, Sarah Grey and Terrence Kaufman (1992). *Language contact, creolization, and genetic linguistics*. Univ of California Press.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The Information Bottleneck Method." In: pp. 368–377. URL: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.6.9199.

Trinh, Trieu H. and Quoc V. Le (2019). *Do Language Models Have Common Sense?* URL: https://openreview.net/forum?id=rkgfWh0qKX.

Trischler, Adam, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman (Aug. 2017). "NewsQA: A Machine Comprehension Dataset." In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics, pp. 191–200. DOI: 10.18653/v1/W17-2623. URL: https://aclanthology.org/W17-2623.

Ufomata, Titi (1999). "Major and minor languages in complex linguistic ecologies: the Nigerian experience." In: *International Journal of Educational Development* 19.4, pp. 315–322. ISSN: 0738-0593. DOI: https://doi.org/10.1016/S0738-0593(99)00031-0. URL: https://www.sciencedirect.com/science/article/pii/S0738059399000310.

Van, Hoang, Zheng Tang, and Mihai Surdeanu (2021). "How May I Help You? Using Neural Text Simplification to Improve Downstream NLP Tasks." In: *EMNLP*.

Volansky, Vered, Noam Ordan, and Shuly Wintner (July 2013). "On the features of translationese." In: *Digital Scholarship in the Humanities* 30.1, pp. 98–118. ISSN: 2055-7671. DOI: 10.1093/llc/fqt031. eprint: https://academic.oup.com/dsh/article-pdf/30/1/98/21521905/fqt031.pdf. URL: https://doi.org/10.1093/llc/fqt031.

Vries, Wietse de, Martijn Wieling, and Malvina Nissim (May 2022). "Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages." In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 7676–7685. DOI: 10.18653/v1/2022.acl-long.529. URL: https://aclanthology.org/2022.acl-long.529.

Wang, Alex and Kyunghyun Cho (2019). "BERT has a Mouth, and It Must Speak: BERT as a Markov Random Field Language Model." In: *arXiv preprint arXiv:1902.04094*.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems." In: *NeurIPS*.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman (Nov. 2018). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://aclanthology.org/W18-5446.

Wang, Bailin, Richard Shin, Xiaodong Liu, Oleksandr Polozov, and Matthew Richardson (July 2020). "RAT-SQL: Relation-Aware Schema Encoding and Linking for Text-to-SQL Parsers." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7567–7578. DOI: 10.18653/v1/2020.acl-main.677. URL: https://aclanthology.org/2020.acl-main.677.

Wang, Hongmin, Jie Yang, and Yue Zhang (May 2019). "From Genesis to Creole Language: Transfer Learning for Singlish Universal Dependencies Parsing and POS Tagging." In: *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19.1. ISSN: 2375-4699. DOI: 10.1145/3321128. URL: https://doi.org/10.1145/3321128.

Wang, Hongmin, Yue Zhang, GuangYong Leonard Chan, Jie Yang, and Hai Leong Chieu (July 2017). "Universal Dependencies Parsing for Colloquial Singaporean English." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1732–1744. DOI: 10.18653/v1/P17-1159. URL: https://aclanthology.org/P17-1159.

Wang, Yongqi (2020). "The Metaphoric and Metonymic Use of Country Names in Economic NewsA Corpus-Based Analysis." In: *Chinese Journal of Applied Linguistics* 43.4, pp. 439 –454. DOI: https://doi.org/10.1515/CJAL-2020-0029. URL: https://www.degruyter.com/view/journals/cjal/43/4/article-p439.xml.

Wang, Yushi, Jonathan Berant, and Percy Liang (July 2015). "Building a Semantic Parser Overnight." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 1332–1342. DOI: 10.3115/v1/P15-1129. URL: https://aclanthology.org/P15-1129.

Welbl, Johannes, Pontus Stenetorp, and Sebastian Riedel (2018). "Constructing Datasets for Multi-hop Reading Comprehension Across Documents." In: *Transactions of the Association for Computational Lin-*

guistics 6, pp. 287–302. DOI: 10.1162/tacl_a_00021. URL: https://aclanthology.org/Q18-1021.

Winford, Donald (1999). *Variation theory: a view from creole continua*. URL: http://hdl.handle.net/10201/1702.

Wolf, Thomas et al. (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *CoRR* abs/1910.03771. arXiv: 1910.03771. URL: http://arxiv.org/abs/1910.03771.

Wong, Yuk Wah and Raymond Mooney (2006). "Learning for semantic parsing with statistical machine translation." In: *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 439–446.

– (2007). "Learning synchronous grammars for semantic parsing with lambda calculus." In: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 960–967.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. (2016). "Google's neural machine translation system: Bridging the gap between human and machine translation." In: *arXiv preprint arXiv:1609.08144*.

Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel (June 2021). "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.

Yakpo, Kofi (2021). "Two types of language contact involving English Creoles: Why Krio (Sierra Leone) has evolved more towards English than its relative Pichi (Equatorial Guinea) towards Spanish." In: *English Today*, 1–12. DOI: 10.1017/S0266078421000146.

Yang, Zhilin, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering." In: *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yin, Pengcheng, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig (2018). "Learning to Mine Aligned Code and Natural Language Pairs from Stack Overflow." In: *International Conference on Mining Software Repositories*. MSR. ACM, pp. 476–486. DOI: https://doi.org/10.1145/3196398.3196408.

Yu, Tao, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong (2020). *GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing*. arXiv: 2009.13845 [cs.CL].

Yu, Tao, Chien-Sheng Wu, Xi Victoria Lin, Bailin Wang, Yi Chern Tan, Xinyi Yang, Dragomir Radev, Richard Socher, and Caiming Xiong (2021). "GraPPa: Grammar-Augmented Pre-Training for Table Semantic Parsing." In: *ArXiv* abs/2009.13845.

Yu, Tao et al. (2018). "Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 3911–3921. DOI: 10.18653/v1/D18-1425. URL: https://aclanthology.org/D18-1425.

Zhao, Jieyu, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang (Sept. 2017). "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2979–2989. DOI: 10.18653/v1/D17-1323. URL: https://www.aclweb.org/anthology/D17-1323.

Zheng, Yinhe, Guanyi Chen, and Minlie Huang (2020). *Out-of-domain Detection for Natural Language Understanding in Dialog Systems*. arXiv: 1909.03862 [cs.CL].

Zhong, Victor, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer (Nov. 2020). "Grounded Adaptation for Zero-shot Executable Semantic Parsing." In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6869–6882. DOI: 10.18653/v1/2020.emnlp-main.558. URL: https://aclanthology.org/2020.emnlp-main.558.

Zhong, Victor, Caiming Xiong, and Richard Socher (2017). "Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning." In: *CoRR* abs/1709.00103.

Zopf, Markus, Maxime Peyrard, and Judith Eckle-Kohler (Dec. 2016). "The Next Step for Multi-Document Summarization: A Heterogeneous Multi-Genre Corpus Built with a Novel Construction Approach." In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. Osaka, Japan: The COLING 2016 Organizing Committee, pp. 1535–1545. URL: https://aclanthology.org/C16-1145.