Overview of the Shared Task on Machine Translation in Dravidian Languages

Anand Kumar Madasamy[1], **Asha Hegde**[2], **Shubhanker Banerjee**[3], **Bharathi Raja Chakravarthi**[3], **Ruba Priyadarshini**[4], **Hosahalli Lakshmaiah Shashirekha**[2] **and John Philip McCrae**[3]

[1]National Institute of Technology Karnataka Surathkal, [2]Mangalore University,
[3]National University of Ireland Galway, [4]Madurai Kamaraj University
m_anandkumar@nitk.edu.in

Abstract

This paper presents an outline of the shared task on translation of under-resourced Dravidian languages at DravidianLangTech-2022 workshop to be held jointly with ACL 2022. A description of the datasets used, approach taken for analysis of submissions and the results have been illustrated in this paper. Five sub-tasks organized as a part of the shared task include the following translation pairs: Kannada to Tamil, Kannada to Telugu, Kannada to Sanskrit, Kannada to Malayalam and Kannada to Tulu. Training, development and test datasets were provided to all participants and results were evaluated on the gold standard datasets. A total of 16 research groups participated in the shared task and a total of 12 submission runs were made for evaluation. Bilingual Evaluation Understudy (BLEU) score was used for evaluation of the translations.

1 Introduction

The results of the shared task on Machine Translation (MT) of Dravidian languages held as a part of DravidianLangTech-2022 workshop have been presented in this paper. Five translation sub-tasks featured in this shared task, namely: Kannada to Tamil, Kannada to Telugu, Kannada to Sanskrit, Kannada to Malayalam and Kannada to Tulu. We evaluated the performance of the systems using BLEU scores. Training, development, and test data used in this shared task have been released publicly. MT is one of the fundamental problems in the area of natural language processing. We hope that this shared task and associated datasets can further research and development of translation technology for under-resourced Dravidian languages.

Related works have been described in section 2. A brief description about Dravidian languages and Sanskrit are given in section 3 and section 4 respectively. The task description and the datasets have been discussed in section 5. The description of the systems submitted has been given to section 6. Lastly, the results and the conclusion have been discussed in section 7 and section 8 respectively.

2 Related Works

In the past few years Deep Learning (DL) based architectures have increasingly been applied to tackle the problem of MT (Pan et al., 2021; Du et al., 2021; Chen et al., 2018; Hoang et al., 2018). These architectures require large amounts of data during training and this, in turn, makes them unsuitable for application in development of translation systems for under-resourced languages. Dabre et al. (2019); Aharoni et al. (2019) demonstrate good performance on translation of under-resourced languages using multilingual MT systems. Another noteworthy approach to tackle this problem is the development of universal translation systems (Gu et al., 2018). The key idea driving this line of research is the development of a system that's capable of transferring linguistic attributes across data from different languages. This is aimed at alleviating the need for large bilingual datasets for under-resourced languages.

Data augmentation is another approach that has been explored in building of translation systems of under-resourced languages. Xia et al. (2019) propose a framework for a translation system that uses monolingual target side dataset along with pivots grounded in a third high resource language. Precisely, they propose a two-stage framework based on pivoting to convert data from high-resourced languages to under-resourced languages, thus augmenting the available data for the translation under-resourced languages.

Another avenue of interest that has been popular amongst researchers working in this domain is application of Transfer Learning (TL) based approaches to improve the performance of MT systems for under-resourced languages. Zoph et al. (2016) train a model for under-resourced MT by initializing some parameters of the model with pa-

Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pages 271 - 278
May 26, 2022 ©2022 Association for Computational Linguistics

rameters from a neural model trained on the task of MT for a resource rich language pair. They report an average increase in performance by 5.6 BLEU. Kocmi and Bojar (2018) demonstrate improved performance on translation of under-resourced languages by employing a simple TL based approach wherein they train a parent model for MT of a resource rich language pair followed by fine-tuning on an under-resourced language pair. It is interesting to note that the authors report improved performance even if the languages in the under-resourced setting are altogether different from the languages which are used to train the model. Mahata et al. (2020) study the impact of languages and their relative position in the language family on the performance of TL systems. Furthermore, they try to quantify the impact of shared vocabulary on the performance of such systems.

In the past few years MT of Indian languages has gained increasing traction from the research community. Chakravarthi et al. (2019, 2021) propose a translation system to improve WordNet for Dravidian languages. Chakravarthi et al. (2019) assess the suitability of using orthographically motivated methods to develop translation systems for Dravidian languages. The key idea behind developing these systems is to leverage the orthographic similarity amongst Dravidian languages to build robust systems in under-resourced scenarios. Pathak and Pakray (2019) propose a neural system for MT of Indian languages based on openNMT[1].

3 Dravidian Languages

Dravidian languages, which make up the fifth largest linguistic family in the world, are spoken by around 200 million people in South Asia and diaspora communities around the world. In Dravidian language family, there are 26 languages, including Tamil, Malayalam, Kannada, and Telugu, which are considered as major languages, in addition to 20 non-literary languages (Krishnamurti, 2003). Since the most Dravidian languages have their writing script, they have a separate block in the Unicode computing industry standard (Sarveswaran et al., 2021). All of these languages use left-to-right writing systems and maintain similar features in their word formation and sentence structure. In these languages, sentences are constructed by a sequence of words and words are formed by adding prefixes and/or suffixes to the root word (Priyadharshini et al., 2021; Kumaresan et al., 2021; Chakravarthi and Muralidaran, 2021; Chakravarthi et al., 2020; Sampath et al., 2022; Ravikiran et al., 2022; Chakravarthi et al., 2022; Bharathi et al., 2022; Priyadharshini et al., 2022). Dravidian languages follow an alpha-syllabic writing scheme, with each character being called a syllable. Consonant ligatures are formed when vowels and consonants are tied together with grammar (Thavareesan and Mahesan, 2019a, 2020a).

Tamil was the first language to be listed as a classical language of India and is one of the longest-surviving classical languages of India. Being a scheduled language by the Indian constitution, it is an official language of Tamil Nadu, a state of India and Puducherry, a territory of India. Further, it is also considered as one of the official languages of Sri Lanka and Singapore. Besides Kerala, Karnataka, Andhra Pradesh, Telangana, and the Union Territory of Andaman and Nicobar Islands, Tamil is spoken by significant minorities in four other south Indian states. Tamil script was first recorded in 580 BCE on pottery from Keezhadi, Sivagangai, and Madurai districts of Tamil Nadu, India by the Tamil Nadu State Department of Archaeology and Archaeological Survey of India (Sivanantham and Seran, 2019). The script was known as Tamili or Tamil-Brahmi[2]. The alphabets of Tamil consist of 18 consonants, 12 vowels, and 216 compound letters followed by a special character making total of 247 letters (Hewavitharana and Fernando, 2002). Tamil is an official language of Tamil Nadu, Sri Lanka, Singapore, and the Union Territory of Puducherry in India. Significant minority speak Tamil in the four other South Indian states of Kerala, Karnataka, Andhra Pradesh, and Telangana, as well as the Union Territory of the Andaman and Nicobar Islands (Sakuntharaj and Mahesan, 2021, 2017, 2016; Thavareesan and Mahesan, 2019b, 2020b,c, 2021). It is also spoken by the Tamil diaspora, which may be found in Malaysia, Myanmar, South Africa, the United Kingdom, the United States, Canada, Australia, and Mauritius. Tamil is also the native language of Sri Lankan Moors. Tamil, one of the 22 scheduled languages in the Indian Constitution, was the first to be designated as a classical language of India (Subalalitha, 2019; Srinivasan and Subalalitha, 2019; Narasimhan et al., 2018). Tamil is one of the world's longest-surviving classical languages. The

[1] https://github.com/OpenNMT/OpenNMT-py

[2] Tamil-Brahmi

earliest epigraphic documents discovered on rock edicts and "hero stones" date from the 6th century BC. Tamil has the oldest ancient non-Sanskritic Indian literature of any Indian language (Anita and Subalalitha, 2019b,a; Subalalitha and Poovammal, 2018).

Malayalam belongs to the Dravidian language family and is highly agglutinative. It originated during the last quarter of the 9th Century A.D (Sekhar, 1951). As a result of the steep Western Ghats separating the dialect from the main speech group in the 16th century, it gradually developed into a separate language. The Ramacaritam is the first literary work written in Malayalam, a combined language of Tamil and Sanskrit, utilizing the Tamil Grantha script used in Tamil Nadu for the writing of Sanskrit and foreign words (Andronov, 1996). There are 13 vowels, 36 consonants, 5 chillu, an anuswara, a visarga, and a chandrakkala making total of 57 letters in Malayalam (Kumar and Chandran, 2015). Telugu belongs to the Dravidian language family and is predominantly spoken by the people of Andra Pradesh. It is the official language of Andhra Pradesh and Telangana with more than 2.75 million Telugu speakers[3]. Inscriptions of Telugu date back to 575 CE. There is a total of 52 letters in Telugu with 16 vowels and 36 consonants and the script is called Abugida which belongs to the Brahmi family[4]. Kannada is the second-oldest Dravidian language, spoken primarily by residents of Karnataka. There are around 44 million Kannada speakers worldwide, with over 12.6 million non-Kannada speakers in Karnataka speaking it as a second or third language[5]. It is one of the scheduled languages of the Indian constitution, as well as the official and administrative language of Karnataka, India. It uses the Brahmi script, which comprises 49 letters in total, comprising 13 vowels, 2 diphthongs, and 34 consonants[6]. Kannada has a large number of articles, although they are not all digitized. Tulu is a prominent Dravidian language spoken primarily by the people of Dakshina Kannada and Udupi in Karnataka state, as well as some parts of Kasaragod in Kerala state. Tulu is spoken by around 2.5 million individuals who believe it to be their mother tongue[7]. With its particular sociocultural qualities, religious practices, creative traditions, and dramatic forms, the Tulu-speaking people have made a substantial contribution to Karnataka's cultural history, and via it, to Indian culture and civilization as a whole. It has kept numerous characteristics of the ancient Dravidian languages while also making some advances not seen in other Dravidian languages (Kekunnaya, 1994). Furthermore, Tulu has its own script, Tigalari, which is developed from the Grantha script, which is no longer in use (Antony et al., 2016). There are 52 letters in Tulu with 16 vowels and 36 consonants.

[3]Telugu language
[4]Teligu-script
[5]Census report 2011
[6]Kannada-script
[7]Tulu language and its script

Languages	Train set	Dev set	Test set
Kannada-Tamil	90,974	2,000	2,000
Kannada-Malayalam	88,813	2,000	2,000
Kannada-Telugu	88,503	2,000	2,000

Table 1: Statistics of set I

Languages	Train set	Dev set	Test set
Kannada-Sanskrit	9,470	1,000	1,000
Kannada-Tulu	8,300	1,000	1,000

Table 2: Statistics of set II

4 Sanskrit

The Sanskrit language has been around for hundreds of years, and it uses the Devanagari (Keith, 1993). With its extensive vocabulary, phonology, grammar, and syntax, Sanskrit literature has a long history of use in ancient poetry, drama, science, and philosophy (Macdonell, 1915). It consists of 16 vowels and 36 consonants and belongs to the Indo-European language family. Sanskrit is a highly inflected language divided into eight chapters to make it more structured and understandable (Panini Asthadhyayi) (Kak, 1987). Despite the enormous number of articles, the quantity of digital resources is limited, especially for the parallel corpus.

5 Task Description and Dataset

Codalab was used to host the shared task. Several translation sub-tasks were organized as a part of this task, namely: Kannada to Tamil, Kannada to Malayalam, Kannada to Telugu, Kannada to Sanskrit, and Kannada to Tulu. The participants could choose which sub-tasks they wanted to participate in. For each language pair, participants were provided with training, development, and test datasets.

Objective of the task was to train/develop MT systems for the language pairs that were provided. Participants translated the test data using MT models proposed by them and submitted the results to the workshop organizers. BLEU is selected as the evaluation metric to evaluate the submitted MT models. In order to determine the participants' rank, the submissions were compared with gold-standard data.

5.1 Dataset

Datasets used in this shared task are broadly grouped into two categories: i) Collection of publicly available parallel corpora (set I) (ii) Construction of parallel corpus from scratch (set II). In the set I, parallel corpora were collected from *Samanantar*[8] - a collection of the largest parallel corpora available for Indic languages (Ramesh et al., 2022) and statistics of set I is shown in Table 1. It may be noted that only a small portion is used in this task instead of using whole dataset. For set II, dataset is manually constructed and Table 2 gives the statistics of set II. Since there is no parallel corpus available for the translation of Kannada-Tulu and Kannada-Sanskrit, the construction of parallel corpora will exacerbate entanglement for these under-resourced language pairs. To create these parallel corpora, we collected monolingual Tulu and Sanskrit documents from digitally accessible sources and manually translated the corresponding Kannada sentences.

6 System Description

Out of 16 research groups, 12 run submissions were made by 4 teams. Set II received the maximum number of submissions (4 teams) followed by set I (3 teams). Further, results of the participated systems in terms of BLEU score and system ranks for each language pair are shown in Table 3. Based on the BLEU scores, we evaluated the performance of the submitted systems. The following is a brief description of the participants' systems. For more information, please refer to their papers.

Aditya et al. (2022) have used two distinct models, namely: i) fine-tuned multilingual indicTrans[9] model with pseudo data generated from monolingual data obtained using backtranslation ii) Convolutional Neural Network (CNN), Seq2Seq models like, Long Short Term Memory (LSTM), Bidirectional LSTM (BiLSTM) and transformer models which were trained from scratch using Fairseq[10] library. They report better BLEU scores for transformer (Vaswani et al., 2017) model trained from scratch using Fairseq library for all the language pairs.

Piyushi et al. (2022) have proposed a system based on the openNMT-py implementation of the transformer (Vaswani et al., 2017) for building the baseline model. Furthermore, they also carry out experiments by using the IndicNLP [11] tokenizer to improve upon the baseline and report an improvement in the observed results. They report better BLEU scores for the Kannada - Tulu and Kannada - Sanskrit languages.

7 Results and Discussion

As shown in Table 3 the submissions were evaluated with BLEU scores. The results indicate that Aditya et al. (2022) achieved the best performance across Kannada - Tamil, Kannada - Telugu and Kannada - Malayalam translation tasks. As mentioned in Section 6, they carried out their experiments with multiple models namely LSTM, BiLSTM, ConvS2S, Transformer, pre-trained multilingual transformer using backtranslation. On these translation tasks they report the better performance of the LSTM based architectures as well as the pre-trained transformer model. This indicates that for these 3 language pairs which have comparatively larger datasets available the DL architectures with a large number of parameters perform better than the other models. For the language pairs in Set II (as shown in 2) the models employed by Aditya et al. (2022) didn't achieve the best performance. The primary reason for this is that size of the dataset for these language pairs is not sufficient to either train the LSTM models from scratch or fine-tune the transformer architecture in order to achieve meaningful generalization.

Piyushi et al. (2022) report the best performance across the Kannada - Tulu and Kannada - Sanskrit language pairs. These languages which belong to Set II (as shown in Table 2) have comparatively smaller datasets. The authors have used openNMT system to tackle the problem at hand. The opotimal performance of their approach for the languages

[8] https://indicnlp.ai4bharat.org/samanantar/

[9] https://indicnlp.ai4bharat.org/indic-trans/

[10] https://github.com/pytorch/fairseq

[11] https://github.com/AI4Bharat/indicnlp_corpus

Languages	Team	BLEU	Rank
Kannada-Tamil	PICT	0.3536	1
	Anvita	0.1791	2
	Translation_Techies	0.0798	3
Kannada-Telugu	PICT	0.3687	1
	Anvita	0.1959	2
	Translation_Techies	0.1242	3
Kannada-Malayalam	PICT	0.2963	1
	Anvita	0.1301	2
	Translation_Techies	0.0729	3
Kannada-Sanskrit	PICT	0.7482	1
	Anvita	0.6209	2
	PICT	0.035	3
	Unitum	0.0011	4
Kannada-Tulu	Translation_Techies	0.6149	1
	Anvita	0.2788	2
	Unitum	0.007	3
	PICT	0.0054	4

Table 3: Results of the participating systems in BLEU score and ranks

of Set II can particularly be attributed to the hyperparameter tuning to the openNMT system. Also, it is interesting to note that participants used the indic tokenization scheme provided by IndicNLP and reported improved results. The impact of the tokenization on specific language pairs however cannot be verified using the subtasks presented in this paper and more comprehensive experiments need to be carried out.

8 Conclusion

The shared task on MT in Dravidian Languages opened up a slew of new research opportunities in the field of MT in Dravidian languages. The task also involves Sanskrit, an ancient language, in addition to Dravidian languages. Despite positive reactions and enthusiasm for attending the event, the number of system submissions was not impressive. We collected Kannada-Tamil, Kannada-Malayalam, and Kannada-Telugu from *samanatar*, a collection of parallel corpora. Further, Kannada-Sanskrit and Kannada-Tulu parallel corpora were created manually. The performance and BLEU scores of the participants are not credible, yet they are not discouraging. The main inference from the participants' results is that along with the baseline MT models, efficient dataset preparation methods, namely, backtranslation and subword tokenization also necessary to achieve better performance in the translation of morphologically rich languages. As a final note, we hope to continue conducting this workshop in the coming years to contribute to the advancement of language technology for under-resourced Dravidian languages.

9 Acknowledgement

Author Bharathi Raja Chakravarthi was supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2 (Insight_2), co-funded by the European Regional Development Fund and Irish Research Council grant IRCLA/2017/129 (CARDAMOM-Comparative Deep Models of Language for Minority and Historical Languages).

Author Shubhanker Banerjee was supported by Science Foundation Ireland under Grant Agreement No. 13/RC/2106_P2 at the ADAPT SFI Research Centre at National University Of Ireland Galway. ADAPT, the SFI Research Centre for AI-Driven Digital Content Technology, is funded by Science Foundation Ireland through the SFI Research Centres Programme.

References

Vyawahare Aditya, Tangsali Rahul, Mandke Aditya, Litake Onkar, and Kadam Dipali. 2022. PICT@DravidianLangTech-ACL2022: Neural Machine Translation on Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*.

Roee Aharoni, Melvin Johnson, and Orhan Firat. 2019. Massively Multilingual Neural Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Mikhail Sergeevich Andronov. 1996. *A Grammar of the Malayalam Language in Historical Treatment*, volume 1. Otto Harrassowitz Verlag.

R Anita and CN Subalalitha. 2019a. An approach to cluster Tamil literatures using discourse connectives. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–4. IEEE.

R Anita and CN Subalalitha. 2019b. Building discourse parser for Thirukkural. In *Proceedings of the 16th International Conference on Natural Language Processing*, pages 18–25.

PJ Antony, CK Savitha, and UJ Ujwal. 2016. Haar features based handwritten character recognition system for Tulu script. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 65–68. IEEE.

B Bharathi, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, N Sripriya, Arunaggiri Pandian, and Swetha Valli. 2022. Findings of the shared task on Speech Recognition for Vulnerable Individuals in Tamil. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Mihael Arcan, and John P. McCrae. 2019. Comparison of Different Orthographies for Machine Translation of Under-Resourced Dravidian Languages. In *2nd Conference on Language, Data and Knowledge (LDK 2019)*, volume 70 of *OpenAccess Series in Informatics (OASIcs)*, pages 6:1–6:14, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Bharathi Raja Chakravarthi and Vigneshwaran Muralidaran. 2021. Findings of the shared task on hope speech detection for equality, diversity, and inclusion. In *Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 61–72, Kyiv. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John Phillip McCrae, Paul Buitaleer, Prasanna Kumar Kumaresan, and Rahul Ponnusamy. 2022. Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Shardul Suryawanshi, Navya Jose, Elizabeth Sherly, and John P McCrae. 2020. Overview of the track on sentiment analysis for Dravidian languages in code-mixed text. In *Forum for Information Retrieval Evaluation*, pages 21–24.

Bharathi Raja Chakravarthi, Priya Rani, Mihael Arcan, and John P McCrae. 2021. A survey of orthographic information in machine translation. *SN Computer Science*, 2(4):1–19.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, et al. 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–86.

Raj Dabre, Atsushi Fujita, and Chenhui Chu. 2019. Exploiting multilingualism through multistage fine-tuning for low-resource neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1410–1416.

Cunxiao Du, Zhaopeng Tu, and Jing Jiang. 2021. Order-agnostic cross entropy for non-autoregressive machine translation. In *International Conference on Machine Learning*, pages 2849–2859. PMLR.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor OK Li. 2018. Universal Neural Machine Translation for Extremely Low Resource Languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354.

S Hewavitharana and HC Fernando. 2002. A Two Stage Classification approach to Tamil Handwriting Recognition. *Tamil Internet*, 2002:118–124.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24.

Subhash C Kak. 1987. The Paninian Approach to Natural Language Processing. *International Journal of Approximate Reasoning*, 1(1):117–130.

Arthur Berriedale Keith. 1993. *A History of Sanskrit Literature*. Motilal Banarsidass Publishe.

K Padmanabha Kekunnaya. 1994. *A Comparative Study of Tulu Dialects*. Rashtrakavi Govinda Pai Research Centre.

Tom Kocmi and Ondrej Bojar. 2018. Trivial Transfer Learning for Low-Resource Neural Machine Translation. *WMT 2018*, page 244.

Bhadriraju Krishnamurti. 2003. *The Dravidian Languages*. Cambridge University Press.

Manoj Kumar and Sandeep Chandran. 2015. Handwritten Malayalam Word Recognition System using Neural Networks. *Int J Eng Res Technol (IJERT)*, 4(4):90–99.

Prasanna Kumar Kumaresan, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Forum for Information Retrieval Evaluation*, pages 16–18.

Arthur Anthony Macdonell. 1915. *A History of Sanskrit Literature*, volume 3. D. Appleton.

Sainik Kumar Mahata, Subhabrata Dutta, Dipankar Das, and Sivaji Bandyopadhyay. 2020. Performance Gain in Low Resource MT with Transfer Learning: An Analysis Concerning Language Families. In *Forum for Information Retrieval Evaluation*, FIRE 2020, page 58–61, New York, NY, USA. Association for Computing Machinery.

Anitha Narasimhan, Aarthy Anandan, Madhan Karky, and CN Subalalitha. 2018. Porul: Option generation and selection and scoring algorithms for a tamil flash card game. *International Journal of Cognitive and Language Sciences*, 12(2):225–228.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive Learning for Many-to-many Multilingual Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 244–258.

Amarnath Pathak and Partha Pakray. 2019. Neural machine translation for Indian languages. *Journal of Intelligent Systems*, 28(3):465–477.

Goyal Piyushi, Supriya Musica, Acharya U Dinesh, and Nayak Ashalatha. 2022. Translation Techies @DravidianLangTech-ACL2022-Machine Translation in Dravidian Languages. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Sajeetha Thavareesan, Dhivya Chinnappa, Durairaj Thenmozhi, and Rahul Ponnusamy. 2021. Overview of the DravidianCodeMix 2021 shared task on sentiment detection in Tamil, Malayalam, and Kannada. In *Forum for Information Retrieval Evaluation*, pages 4–6.

Gowtham Ramesh, Sumanth Doddapaneni, Aravinth Bheemaraj, Mayank Jobanputra, Raghavan AK, Ajitesh Sharma, Sujit Sahoo, Harshita Diddee, Divyanshu Kakwani, Navneet Kumar, et al. 2022. Samanantar: The Largest Publicly Available Parallel Corpora Collection for 11 Indic Languages. *Transactions of the Association for Computational Linguistics*, 10:145–162.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the shared task on Offensive Span Identification in code-mixed Tamil-English comments. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2016. A novel hybrid approach to detect and correct spelling in Tamil text. In *2016 IEEE International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 1–6.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2017. Use of a novel hash-table for speeding-up suggestions for misspelt Tamil words. In *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, pages 1–5.

Ratnasingam Sakuntharaj and Sinnathamby Mahesan. 2021. Missing word detection and correction based on context of Tamil sentences using n-grams. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 42–47.

Anbukkarasi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, Ruba Priyadharshini, Subalalitha Chinnaudayar Navaneethakrishnan, Kogilavani Shanmugavadivel, Sajeetha Thavareesan, Sathiyaraj Thangasamy, Parameswari Krishnamurthy, Adeep Hande, Sean Benhur, Kishor Kumar Ponnusamy, and Santhiya Pandiyan. 2022. Findings of the shared task on Emotion Analysis in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Kengatharaiyer Sarveswaran, Gihan Dias, and Miriam Butt. 2021. Thamizhimorph: A Morphological

Parser for the Tamil Language. *Machine Translation*, 35(1):37–70.

A. C. Sekhar. 1951. Evolution of Malayalam. *Bulletin of the Deccan College Research Institute*, 12(1/2):1–216.

R Sivanantham and M Seran. 2019. Keeladi: An Urban Settlement of Sangam Age on the Banks of River Vaigai. *India: Department of Archaeology, Government of Tamil Nadu, Chennai*.

R Srinivasan and CN Subalalitha. 2019. Automated named entity recognition from tamil documents. In *2019 IEEE 1st International Conference on Energy, Systems and Information Processing (ICESIP)*, pages 1–5. IEEE.

C. N. Subalalitha. 2019. Information extraction framework for Kurunthogai. *Sādhanā*, 44(7):156.

CN Subalalitha and E Poovammal. 2018. Automatic bilingual dictionary construction for Tirukural. *Applied Artificial Intelligence*, 32(6):558–567.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019a. Sentiment analysis in Tamil texts: A Study on Machine Learning Techniques and Feature Representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2019b. Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. In *2019 14th Conference on Industrial and Information Systems (ICIIS)*, pages 320–325.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020a. Sentiment Lexicon Expansion using Word2vec and fastText for Sentiment Prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276. IEEE.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020b. Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts. In *2020 Moratuwa Engineering Research Conference (MERCon)*, pages 272–276.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2020c. Word embedding-based part of speech tagging in Tamil texts. In *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*, pages 478–482.

Sajeetha Thavareesan and Sinnathamby Mahesan. 2021. Sentiment analysis in Tamil texts using k-means and k-nearest neighbour. In *2021 10th International Conference on Information and Automation for Sustainability (ICIAfS)*, pages 48–53.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Advances in neural information processing systems*, 30.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized Data Augmentation for Low-Resource Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575.