TSD: Towards Computational Processing of Tamil Similes – A Tamil Simile Dataset

Aathavan Nithiyananthan, Jathushan Raveendra, Uthayasanker Thayasivam
Department of Computer Science and Engineering, University of Moratuwa,
Sri Lanka
{aathavan.20, jathushan.20, rtuthaya}@cse.mrt.ac.lk

Abstract

A simile is a powerful figure of speech that makes a comparison between two different things via shared properties, often using words like "like" or "as" to create vivid imagery, convey emotions, and enhance understanding. However, computational research on similes is limited in low-resource languages like Tamil due to the lack of simile datasets. This work introduces a manually annotated Tamil Simile Dataset (TSD) comprising around 1.5k simile sentences drawn from various sources. Our data annotation guidelines ensure that all the simile sentences are annotated with the three components, namely *tenor*, *vehicle*, and *context*. We benchmark our dataset for simile interpretation and simile generation tasks using chosen pre-trained language models (PLMs) and present the results. Our findings highlight the challenges of simile tasks in Tamil, suggesting areas for further improvement. We believe that TSD will drive progress in computational simile processing for Tamil and other low-resource languages, further advancing simile related tasks in Natural Language Processing.

1 Introduction

A simile is a figure of speech that explicitly compares two different things by saying that one thing is like another, so it typically contains comparison expressions such as "like" and "as" (Paul, 1970). Similes allow people to create vivid images and convey emotions in ways that literal language cannot. Computational processing of similes is gaining attention in Natural Language Processing (NLP) research which enables the development of more engaging conversational systems (Zheng et al., 2020), creative writing tools (Zhang et al., 2021) and also enhances applications in sentiment analysis (Ge et al., 2023).

Research on simile processing is limited compared to other areas in NLP (Chakrabarty et al., 2022). Early research lacked dedicated datasets which made many researchers to create small, task-specific ones as the field evolved (Ge et al., 2023). The simile identification task has achieved significant advances, but other tasks have yet to gain more traction due to the lack of specifically designed annotated data (Ge et al., 2023). There is still room for the development of theories that abstractly explain the connection between the compared elements in similes (Lai and Nissim, 2024). Due to these constraints, simile related tasks are still challenging for high-resourced languages like English and Chinese in which most of the research in similes is centered.

[இமை]$_{\text{VEHICLE}}$ போலக் [காக்கிறான்]$_{\text{CONTEXT}}$ [கடவுள்]$_{\text{TENOR}}$. [God]$_{\text{TENOR}}$ [protects]$_{\text{CONTEXT}}$ like an [eyelash]$_{\text{VEHICLE}}$.
அவள் [உடம்பு]$_{\text{TENOR}}$ காற்றில் ஆடிய [மரங்களைப்]$_{\text{VEHICLE}}$ போல் [ஆடியது]$_{\text{CONTEXT}}$. Her [body]$_{\text{TENOR}}$ [swayed]$_{\text{CONTEXT}}$ like [trees]$_{\text{VEHICLE}}$ dancing in the wind.
[அலைகள்]$_{\text{VEHICLE}}$ போலவே [மோதும்]$_{\text{CONTEXT}}$ உந்தன் [ஞாபகம்]$_{\text{TENOR}}$. Your [memory]$_{\text{TENOR}}$ [strikes]$_{\text{CONTEXT}}$ like [waves]$_{\text{VEHICLE}}$.

Table 1: Examples of Tamil simile sentences and components.

Considering low-resourced languages, research in computational simile processing is extremely limited. There are no large datasets or established resources, and only a few studies have focused on figurative language like similes in these languages. Tamil is a language with unique cultural and linguistic expressions due to its agglutinative nature and complex morphological structures (Keane, 2004). Languages like Tamil are often overlooked in NLP research due to the lack of sufficient resources like annotated datasets and tools.

Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 10–16

The creation of simile datasets has improved over time with different labeling methods as research and tasks developed (Ge et al., 2023). Different simile tasks require datasets with specifically annotated components of similes. Currently, it has become standard to annotate all the components, as this can be used across tasks (Yang et al., 2023; Shao et al., 2024a).

In this work, we present Tamil Simile Datset (TSD), a simile dataset annotated manually with all the components of the simile. The contributions of this paper are:

1. We present the Tamil Simile Dataset (TSD), which is the first simile dataset for the Tamil language.

2. Our monolingual dataset contains 1520 sentences all annotated with TENOR, VEHICLE, and CONTEXT.

3. We evaluate our dataset for Simile interpretation and Simile Generation tasks using chosen pre-trained language models and present the results.

2 Background

A simile (உவமையணி) is a figure of speech in which one concept is described in terms of another known concept that shares similar properties, typically using comparators like "like" (போல) to emphasize the comparison. In a simile, the word or concept which is being described is the TENOR (உவமேயம்). The word or concept used to describe the TENOR is the VEHICLE (உவமானம்). VEHICLE is a component that brings imagery or qualities to mind for comparison. Additionally, CONTEXT (பொதுத்தன்மை) is the property through which the comparison is made. Examples of simile sentences and annotated components are shown in Table 1.

3 Related Works

3.1 Simile Datasets

Several datasets have been developed in English and Chinese, which are high-resourced languages (Joshi et al., 2020) to support research in simile-related tasks. Self Labeled Simile (SLS) dataset (Chakrabarty et al., 2020) and the Writing Polishment Similes (WPS) dataset (Zhang et al., 2021) consists of automatically annotated sentences and were used for simile generation tasks. Chinese Metaphor (CM) dataset (Su et al., 2016), CMC dataset (Li et al., 2022), and MSD dataset (Ma et al., 2023) were annotated with TENOR and VEHICLE. These datasets were used for interpretation and generation tasks. MCP dataset (He et al., 2022), GraCe dataset (Yang et al., 2023), and the most recent CMDAG dataset (Shao et al., 2024a) are annotated with all three components and are also utilized for both simile interpretation and simile generation tasks.

Research on similes remains limited in Dravidian languages (Paul et al., 2024). In low-resource languages, small-scale efforts exist, such as simile generation in Afrikaans (van Heerden and Bas, 2021), and recently, a Malayalam simile dataset for identification has been developed (Paul et al., 2024). Elanchezhiyan et al. (2014) analyzed Tamil song lyrics to identify simile patterns and attributes for automatic simile generation, but no dedicated dataset was released, and further details are unavailable.

3.2 Tasks in Simile Processing

Simile interpretation and simile generation are the two main directions of the simile study (Yu and Wan, 2019).

Simile interpretation task focuses on identifying shared properties between the TENOR and VEHICLE. Early methods relied on word embeddings to measure semantic similarity (Zheng et al., 2020; Bar et al., 2022), but recent work has integrated knowledge bases like ConceptNet (Gero and Chilton, 2019; Stowe et al., 2021). PLMs have further refined interpretation by capturing implicit meanings without predefined rules (Su et al., 2017). Ma et al. (2023) introduced a task where models predict the shared property in a simile, while Chen et al. (2022) used masked language modeling (MLM) to predict missing simile elements.

Simile generation task involves constructing simile expressions. Recent approaches fine-tune pre-trained language models such as GPT-2 (Li et al., 2022) or BART (Lewis, 2019) for this task. Knowledge-driven methods frame it as knowledge graph completion, generating VEHICLEs based on relational context (Song et al., 2020). Chen et al. (2022) refined PLM-based simile matching, while Yang et al. (2023) used CBART (Shao et al., 2024b) with multiple constraints for Chinese simile generation. Ma et al. (2023) extended the task to dialogue systems, requiring models to select

appropriate VEHICLEs. Recent research underscores the importance of PLMs in improving both simile interpretation and generation tasks.

4 Tamil Simile Dataset

In this section, we present the collection, annotation, and statistics of our manually annotated Tamil Simile Dataset.

4.1 Data collection

We collected data from various sources. We used Wikisource API[1] to get random articles from Tamil Wikisource and extracted the article contents which contained Tamil simile comparators such as "போல" (Pola "like") or "போன்ற" (Pondra "like"'). Additionally, we extracted texts which contained morphemes of "போல" such as "போல்" (Pol), "போலே" (Pole) and "போலும்" (Polum)—all of which convey the meaning "like". This included similes from various kinds of literature, such as old Tamil scripts like Kambaramayanam and Tamil poems, stories, and essays. We also extracted similes from Tamil song lyrics, as similes are most frequently used in Tamil songs. We collected songs from tamil2lyrics.com[2] and extracted songs that contained simile comparators as above.

4.2 Data annotation

We employed 10 annotators to extract meaningful sentences from the collected data. In the Tamil language, not all the sentences that contain the comparator "போல" are similes. For example, consider the sentence "கதவு அடைத்து உட்புறமாகத் தாழிட்டிருப்பது போல தெரிந்தது" ("Kathavu adaittu utpuramaga thazhittiruppathu pola therinthathu") (translates to: "The door seemed to be locked and slammed inwards"). Here "போல" is used to convey a state of appearance rather than a direct comparison. So our annotators first extract sentences that are similes and disregard literal sentences.

Sentences annotated as similes are forwarded to the next stage of the annotation. In this stage, another 4 annotators annotated the VEHICLE, TENOR, and the CONTEXT of the sentences. When the annotators extract a sentence as a simile sentence, it will have the comparator and the VEHICLE word by default. So, annotators are asked to annotate the VEHICLE (it can be word/phrase/sentence) first. The next step is to annotate the TENOR and CONTEXT if it is found in the sentence. If not, we instructed the annotators to annotate the TENOR and CONTEXT (both can be a word/phrase/sentence). We asked the annotators to disregard confusing cases where it is difficult to find TENOR or CONTEXT. In this way, we were able to ensure all the simile sentences in our dataset were annotated with all three components. Our annotation process is shown in Figure 1. Before the components annotation phase, a training session was conducted, and annotators were trained with examples and instructed with Tamil simile principles. A set of 50 sample sentences sourced from various genres was given to all 4 annotators for component annotation to check the reliability of their annotation. We computed the inner-annotator agreement of simile component annotation via Krippendorff's alpha (Krippendorff, 2011). The overall agreement rate was found to be 0.78. Statistics of TSD are shown in Table 2.

Measurement	Value
# Simile Sentences	1520
# Distinct Tenors	706
# Distinct Vehicles	1042
# Distinct Contexts	1077
Average # Words per Sentence	6

Table 2: Statistics of the dataset.

5 Tasks

In this section, we introduce 2 tasks for our Tamil simile dataset, including the definition of the tasks, the baselines, evaluation metrics, experimental results, and analysis.

5.1 Simile Interpretation/Generation Tasks

Following prior work on simile interpretation (Song et al., 2020; Zheng et al., 2020; He et al., 2022; Chen et al., 2022; Shuhan et al., 2023) and simile generation (Song et al., 2020; Chen et al., 2022; Shuhan et al., 2023), we define Simile Interpretation/Generation (SI/SG) as a fill-mask objective task. We evaluate the models on 100 samples from our dataset.

For the simile interpretation task, we remove the CONTEXT from the simile sentence and replace it with a blank. The model is required to generate the missing CONTEXT. Similarly, for the simile generation task, we remove the VEHICLE from the

[1] https://ta.wikisource.org/w/api.php
[2] https://www.tamil2lyrics.com/

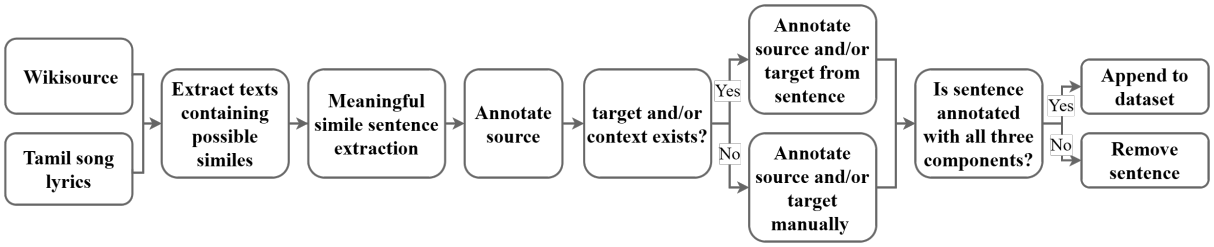

Figure 1: The data annotation process.

Model	Interpretation			Generation		
	MRR↑	Hits@5↑	Hits@10↑	MRR↑	Hits@5↑	Hits@10↑
TamilBERT	0.026	0.05	**0.07**	**0.200**	**0.27**	**0.38**
IndicBERT v2	**0.033**	**0.06**	**0.07**	0.122	0.20	0.30
MuRIL	0.027	**0.06**	0.06	0.140	0.21	0.24
XLM-RoBERTa	0.003	0.01	0.01	0.030	0.05	0.06

Table 3: Simile interpretation and generation results (MRR, Hits@5, Hits@10).

simile sentence, leaving a blank, and the model needs to generate an appropriate VEHICLE. In both cases, we extract the top 10 predicted words from the model.

We fine-tune TamilBERT (Joshi, 2022), IndicBERT v2 (Doddapaneni et al., 2023), MuRIL (Khanuja et al., 2021), and XLM-RoBERTa (Conneau et al., 2019) on the Tamil Simile Dataset. These baselines are chosen due to their strong performance in Dravidian and multilingual NLP tasks, particularly in low-resource settings.

The performance of the models is evaluated using Mean Reciprocal Rank (MRR), Hits@5, and Hits@10. MRR measures the average of the reciprocal ranks of the first correct prediction, providing insight into how well models rank the correct completion. Hits@5 and Hits@10 measure the proportion of cases where the correct word appears within the top 5 and top 10 predictions.

6 Results and Discussion

Table 3 presents the results of simile interpretation and generation tasks. Simile interpretation task yielded lower results compared to the generation task, which may be attributed to the structural characteristics of Tamil simile sentences, where contextual information is sometimes omitted. This aspect requires further investigation. The TamilBERT model achieved relatively high scores in the simile generation task, indicating that a monolingual model trained specifically on Tamil data can be more effective for simile processing in the Tamil language. Additionally, models pre-trained on Indian languages, such as IndicBERT v2 and MuRIL, demonstrated reasonable performance in simile generation. In contrast, XLM-RoBERTa, a multilingual model trained on 100 languages, exhibited weaker performance in simile-related tasks. These findings highlight the impact of language-specific pretraining in low-resource NLP, particularly for complex tasks like simile processing.

Interestingly, the simile interpretation task showed significant improvement during the fine-tuning phase. Initially, the models generated irrelevant tokens such as “##ாதே”, “-”, and “"##ுது”. After fine-tuning, the predictions were contextually appropriate, including words like “பறக்கும்” (flying), “வளைந்த” (curved), and “அழகான” (beautiful). However, we found that many error predictions occurred when the CONTEXT was not a noun or when morphemes complicated interpretation. Further investigation into the effect of tokenization on this task could provide deeper insights into these behaviors. For the generation task, models struggled to predict words that are not so commonly occurring in Tamil language. Exploring alternative fine-tuning techniques may improve the model’s ability to generate more relevant predictions.

Our results and findings indicate that both simile interpretation and generation are challenging for the Tamil language. This can be attributed to Tamil’s linguistic complexities, which make these tasks more difficult compared to languages

with simpler structures. These challenges present valuable opportunities for future research, and the Tamil Simile Dataset (TSD) can serve as a valuable resource for advancing simile processing in low-resource languages.

7 Conclusion

We present a manually annotated Tamil simile dataset (TSD) comprising 1520 simile sentences sourced from a wide range of Tamil literary forms, including poems, short stories, articles, and song lyrics. Our dataset annotators achieved inter-annotator agreement of 0.78, underscoring the reliability of our dataset. We also benchmark our dataset for simile interpretation and simile generation tasks using pre-trained language models. Our results show that simile-related tasks are challenging for Tamil Language. This shows that our dataset has great potential to help improve the understanding and creation of Tamil similes.

8 Limitations

When annotators annotate components TENOR and/or CONTEXT that are not in the original simile sentence manually, there is a possibility of multiple suitable TENORs and/or CONTEXTs for that simile. However, in our dataset, the appropriate one, as determined by the annotators is annotated. The 100 examples we used are sentences that had VEHICLE and CONTEXT within them. Sentences from our dataset which are sourced from tamil2lyrics.com, comprise song lyrics from the 1950s to 2023. This covers a wide range of timelines and songs, though not every song is included. In addition, our dataset consists of different Tamil literary forms such as poems, articles, and other literary sentences extracted from Wikisource. Our dataset is limited in terms of coverage as we could only get sentences from the extracted pages returned by Wikisource API. While Tamil has been rich in figurative language since ancient times, its usage has evolved over time. Expanding simile datasets to include more classical and historical Tamil literature would enhance coverage and further improve computational simile processing in Tamil.

Acknowledgments

This research was funded by the University of Moratuwa Senate Research Committee (SRC) grant SRC/ST/2024/44.

References

Kfir Bar, Nachum Dershowitz, and Lena Dankin. 2022. Metaphor interpretation using word embeddings. *Computación y Sistemas*, 26(3):1301–1311.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022. It's not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Smaranda Muresan, and Nanyun Peng. 2020. Generating similes effortlessly like a pro: A style transfer approach for simile generation. *arXiv preprint arXiv:2009.08942*.

Weijie Chen, Yongzhu Chang, Rongsheng Zhang, Jiashu Pu, Guandan Chen, Le Zhang, Yadong Xi, Yijiang Chen, and Chang Su. 2022. Probing simile knowledge from pre-trained language models. *arXiv preprint arXiv:2204.12807*.

Alexis Conneau, 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. *arXiv preprint arXiv:1911.02116*.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

K. Elanchezhiyan, E. Tamil Selvi, N. Revathi, G. P. Shanthi, S. Shireen, and Madhan Karky. 2014. Simile generation. In *Proceedings of the 13th International Tamil Internet Conference*.

Mengshi Ge, Rui Mao, and Erik Cambria. 2023. A survey on computational metaphor processing techniques: From identification, interpretation, generation to application. *Artificial Intelligence Review*, 56(Suppl 2):1829–1895.

Katy Ilonka Gero and Lydia B Chilton. 2019. Metaphoria: An algorithmic companion for metaphor creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12.

Qianyu He, Sijie Cheng, Zhixu Li, Rui Xie, and Yanghua Xiao. 2022. Can pre-trained language models interpret similes as smart as human? *arXiv preprint arXiv:2203.08452*.

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

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Elinor Keane. 2004. Tamil. *Journal of the International Phonetic Association*, 34(1):111–116.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha P. Talukdar. 2021. Muril: Multilingual representations for indian languages. *CoRR*, abs/2103.10730.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Huiyuan Lai and Malvina Nissim. 2024. A survey on automatic generation of figurative language: From rule-based systems to large language models. *ACM Computing Surveys*, 56(10):1–34.

M Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yucheng Li, Chenghua Lin, and Frank Guerin. 2022. CM-Gen: A Neural Framework for Chinese Metaphor Generation with Explicit Context Modelling. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6468–6479, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Longxuan Ma, Weinan Zhang, Shuhan Zhou, Churui Sun, Changxin Ke, and Ting Liu. 2023. I run as fast as a rabbit, can you? A Multilingual Simile Dialogue Dataset. *arXiv preprint*. ArXiv:2306.05672 [cs].

Anthony M Paul. 1970. Figurative language. *Philosophy & Rhetoric*, pages 225–248.

Reenu Paul, Wincy Abraham, and Anitha S Pillai. 2024. Malupama-figurative language identification in malayalam-an experimental study. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 357–367.

Yujie Shao, Xinrong Yao, Xingwei Qu, Chenghua Lin, Shi Wang, Stephen W Huang, Ge Zhang, and Jie Fu. 2024a. Cmdag: A chinese metaphor dataset with annotated grounds as cot for boosting metaphor generation. *arXiv preprint arXiv:2402.13145*.

Yunfan Shao, Zhichao Geng, Yitao Liu, Junqi Dai, Hang Yan, Fei Yang, Zhe Li, Hujun Bao, and Xipeng Qiu. 2024b. Cpt: A pre-trained unbalanced transformer for both chinese language understanding and generation. *Science China Information Sciences*, 67(5):152102.

Zhou Shuhan, Ma Longxuan, and Shao Yanqiu. 2023. Exploring Accurate and Generic Simile Knowledge from Pre-trained Language Models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 918–929, Harbin, China. Chinese Information Processing Society of China.

Wei Song, Jingjin Guo, Ruiji Fu, Ting Liu, and Lizhen Liu. 2020. A Knowledge Graph Embedding Approach for Metaphor Processing. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1–1.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. *arXiv preprint arXiv:2106.01228*.

Chang Su, Shuman Huang, and Yijiang Chen. 2017. Automatic detection and interpretation of nominal metaphor based on the theory of meaning. *Neurocomputing*, 219:300–311.

Chang Su, Jia Tian, and Yijiang Chen. 2016. Latent semantic similarity based interpretation of chinese metaphors. *Engineering Applications of Artificial Intelligence*, 48:188–203.

Imke van Heerden and Anil Bas. 2021. Towards figurative language generation in afrikaans. In *Proceedings of the SIGTYP 2021 Workshop on Typology for Cross-Linguistic NLP*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023. Fantastic Expressions and Where to Find Them: Chinese Simile Generation with Multiple Constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486, Toronto, Canada. Association for Computational Linguistics.

Zhiwei Yu and Xiaojun Wan. 2019. How to avoid sentences spelling boring? towards a neural approach to unsupervised metaphor generation. 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 861–871.

Jiayi Zhang, Zhi Cui, Xiaoqiang Xia, Yalong Guo, Yanran Li, Chen Wei, and Jianwei Cui. 2021. Writing polishment with simile: Task, dataset and a neural approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14383–14392.

Danning Zheng, Ruihua Song, Tianran Hu, Hao Fu, and Jin Zhou. 2020. "love is as complex as math": Metaphor generation system for social chatbot. In *Chinese Lexical Semantics: 20th Workshop, CLSW 2019, Beijing, China, June 28–30, 2019, Revised Selected Papers 20*, pages 337–347. Springer.

A Dataset

A.1 Tamil Simile Dataset (TSD)

Examples in TSD are shown in Table 4.

Sentence	Tenor	Vehicle	Context
கடல் போல பெரிதாக நீ நின்றாய்.	கடல்	நீ	பெரிதாக
You stood as big as the sea.	sea	you	big
பறவை போலே பறந்து செல்வோம்.	பறவை	நாம்	பறந்து
Let's fly like a bird.	bird	Let's	fly
ஒரு கோயில் போல் இந்த மாளிகை.	கோயில்	மாளிகை	புனிதமானது
This mansion is like a temple.	temple	mansion	sacred
வழியிலே தங்கத்தகடு போல மின்னிய தவளை தத்திச் சென்றது.	தங்கத்தகடு	தவளை	மின்னிய
On the way, a frog that glittered like a gold plate jumped away.	gold plate	frog	glittered

Table 4: Examples of annotated similes in the TSD.