Studying Contribution Types of Research in Natural Language Processing

Krirk Nirunwiroj **Niyati Bafna**
Johns Hopkins University
{knirunw1,nbafna1}@jh.edu

1 Introduction

Computation Linguistics (CL) and Natural Language Processing (NLP) made fast progress over the last decade, with advances in hardware computing and GPU capacities, data availability with the rise of the internet, and neural network architectures with hugely successful Transformer-based architectures. We are interested in studying the evolution of topics of focus as well as techniques in the field over this time period, using conference papers published by the Association for Computation Linguistics (ACL). ACL is an organization and conference that is one of the top venues for research in CL / NLP.[1]

A research paper in NLP may make an advance to the field in several ways. Its innovation may arise from novel algorithmic/mathematical approaches to problems, from insights on data or models, analytical insights on existing models, or from engineering or architectural innovations to models. Note that these types are orthogonal to the actual problem or subfield being studied, e.g. computational social science, multilinguality, speech technologies, etc. We are mainly interested in the following question: **How has the distribution over the above types of contributions changed over time?**

The NLP research community (like most research communities) is self-guided; i.e. there is no overarching organization who decides what the community should study, and how much effort we should invest in different things. Individual researchers and labs make decisions on relevant and impactful problems to work on based on several factors such as funding, their peers, and personal interests. Given this, it's important to present the large-picture perspective on the evolution of the field to researchers, so that we can step back and assess whether we are distributing our efforts in a reasonable manner over directions of research.

The full code used in this study is publicly available at https://github.com/krirk-n/Contribution-Types-ACL.

2 Data

We use the ACL-OCL dataset,[2] a collection of 73,285 papers from the ACL Anthology spanning from 1952 to 2022. Each entry includes rich metadata such as title, authors, publication year, and venue (`booktitle`), along with textual content fields including `abstract` and `full_text`. A complete list of available fields is included in Appendix A.

Summary Statistics of the Data We restrict our analysis to papers that contain both `abstract` and `full_text` fields, resulting in a cleaned subset of **63,903** papers.

We begin with basic descriptive statistics:

- **Temporal Coverage:** The cleaned data spans publications from 1952 to 2022, with the number of papers per year increasing significantly in the last two decades (see Appendix 4).
- **Venue Distribution:** The top venues include ACL, EMNLP, LREC, and COLING (Appendix 5).

We also conduct an initial LDA topic modeling analysis using 10 topics trained on the abstracts. While a few topics appear interpretable (e.g., syntax, machine translation, dialogue systems), others are vague or overlapping (e.g., multiple topics emphasize generic terms like "model" and "data"), and

[1] It was founded in the 1960's as the Association for Machine Translation and Computational Linguistics (AMTCL) (`https://en.wikipedia.org/wiki/Association_for_Computational_Linguistics`), but dropped the mention of machine translation because the task was deemed "too difficult" (as per an anecdote).

[2] `https://huggingface.co/datasets/WINGNUS/ACL-OCL`

Topic ID	Top Keywords
0	model, neural, models, information, propose, network, sentiment, task, based, attention
1	parsing, based, paper, task, dependency, parser, grammar, features, syntactic, tree
2	models, model, data, training, language, performance, learning, tasks, trained, domain
3	la, des, et, les, en, une, le, nous, dans, pour
4	language, corpus, annotation, data, paper, languages, research, text, present, annotated
5	semantic, paper, discourse, language, structure, linguistic, information, lexical, syntactic, analysis
6	translation, machine, english, based, mt, language, systems, question, paper, quality
7	dialogue, user, human, language, systems, users, speech, based, paper, dialog
8	task, text, evaluation, news, dataset, results, summarization, paper, information, human
9	word, words, semantic, based, method, using, approach, data, results, embeddings

Table 1: Top words from each of the 10 LDA topics trained on abstracts.

one consists primarily of non-English stopwords. This motivates our decision to use neural topic models for improved coherence and robustness. Full topic keywords are shown in Table 1.

Filtering and subsampling We sample 100 papers per year from the most recent 10-year period in the dataset (2013 to 2022, inclusive), resulting in 1,000 papers per condition under three distinct filtering settings: (1) inclusive of all workshops and sub-conferences (*all*), (2) limited only to main conference papers from Empirical Methods in Natural Language Processing (EMNLP), Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL), and ACL (*main*), and (3) papers in the subfield of machine translation (*MT*).
We use `acl-id` for identifying main conference papers, and based on string matches with the term "machine translation" in the title and abstract for identifying machine translation papers; this is likely to have a low false positive rate. See the resulting numbers of papers in Table 2. Note that we only sample 100 papers per year from all filtered subsets.

3 Method

3.1 Annotation Schema

We sampled fifty papers over the last decade and manually characterized their contribution types, with the goal of developing a taxonomy that makes broad sense over time, as particular topics of interest have shifted.

Filtering	#Papers	Year	%Coverage
all	63,903	1952–2022	100.0%
main	10,027	1979–2022	15.7%
MT	7,680	1957–2022	12.0%

Table 2: Number of papers, year range, and relative coverage for each filtering setting.

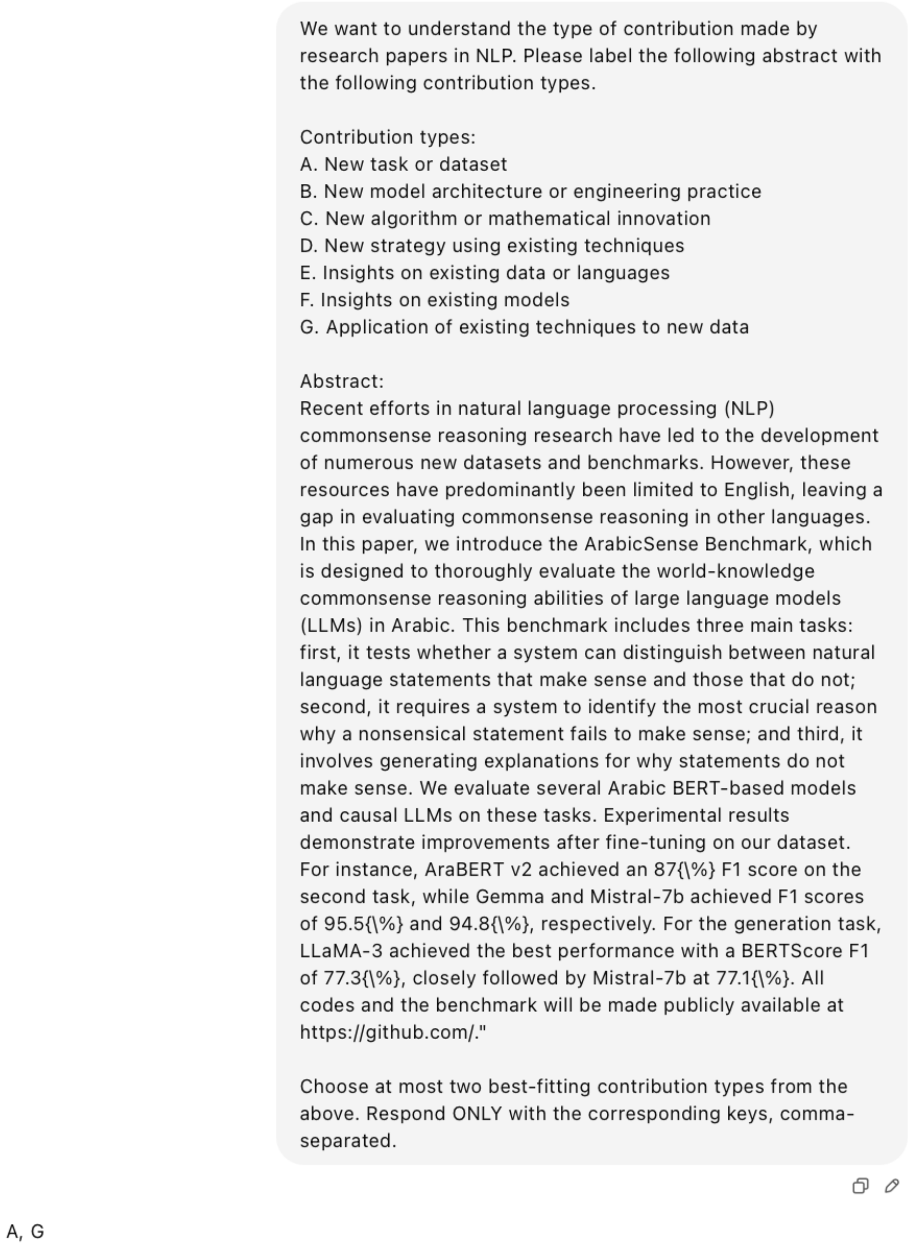
We use the following seven categories for contribution type over the dataset:

- A. New task or dataset
- B. New model architecture or engineering practice
- C. New algorithm or mathematical innovation
- D. New strategy using existing techniques
- E. Insights on existing data or languages
- F. Insights on existing models
- G. Application of existing techniques to new data

See an example of each in Table 3.

3.2 Labeling dataset

We use GPT-4o to perform our annotation. We provide the model with the paper abstract and prompt it to respond with the top two categories applying to the paper using the above scheme. We set $temperature = 0$ for inference and $seed = 42$ for sampling to maintain high reproducibility and determinism. We experimented with various prompts for best performance, and finally went with the prompt as shown in Figure 1. While some of these decisions are somewhat subjective, with natural ambiguity with regard to certain papers, we find from eyeballing the resulting annotations that the model is able to provide reasonable labels based on the content of the abstract. Also note that while providing some in-context examples would probably improve the annotation quality, we decide to go ahead with zero-shot labeling for the following reasons: a) Initial small-scale runs indicate zero-shot labeling is already good enough. b) Providing examples of all labels (probably necessary to prevent bias from the prompt) would increase the context length and annotation cost 6-7 fold.



We want to understand the type of contribution made by research papers in NLP. Please label the following abstract with the following contribution types.

Contribution types:
A. New task or dataset
B. New model architecture or engineering practice
C. New algorithm or mathematical innovation
D. New strategy using existing techniques
E. Insights on existing data or languages
F. Insights on existing models
G. Application of existing techniques to new data

Abstract:
Recent efforts in natural language processing (NLP) commonsense reasoning research have led to the development of numerous new datasets and benchmarks. However, these resources have predominantly been limited to English, leaving a gap in evaluating commonsense reasoning in other languages. In this paper, we introduce the ArabicSense Benchmark, which is designed to thoroughly evaluate the world-knowledge commonsense reasoning abilities of large language models (LLMs) in Arabic. This benchmark includes three main tasks: first, it tests whether a system can distinguish between natural language statements that make sense and those that do not; second, it requires a system to identify the most crucial reason why a nonsensical statement fails to make sense; and third, it involves generating explanations for why statements do not make sense. We evaluate several Arabic BERT-based models and causal LLMs on these tasks. Experimental results demonstrate improvements after fine-tuning on our dataset. For instance, AraBERT v2 achieved an 87{\%} F1 score on the second task, while Gemma and Mistral-7b achieved F1 scores of 95.5{\%} and 94.8{\%}, respectively. For the generation task, LLaMA-3 achieved the best performance with a BERTScore F1 of 77.3{\%}, closely followed by Mistral-7b at 77.1{\%}. All codes and the benchmark will be made publicly available at https://github.com/."

Choose at most two best-fitting contribution types from the above. Respond ONLY with the corresponding keys, comma-separated.

A, G

Figure 1: Prompt and response example

Contribution Type	Example Paper
A. New task or dataset	Khanuja et al. (2024): Creating a new task dataset focusing on cultural adaptation of images.
B. New model architecture or engineering practice	Devlin et al. (2019): Introducing BERT!
C. New algorithm or mathematical innovation	Zheng et al. (2021): Increasing tokenizer vocabulary in an efficient way with KNN sampling.
D. New strategy using existing techniques	Dimakis et al. (2024): Applying in-context learning techniques to boost low-resource performance.
E. Insights on existing data or languages	Melgarejo et al. (2022): (Among other things) Compares Quechua to Spanish using WordNets.
F. Insights on existing models	Robinson et al. (2023): Evaluating ChatGPT for low-resource machine translation.
G. Application of existing techniques to new data	Lin and Su (2021)

Table 3: Our annotation schema, with associated examples.

4 Results and Discussion

We present the overall distribution and temporal evolution of contribution types across three filtering settings: *all*, *main*, and *MT*.

4.1 Distribution Across Contribution Types

Figure 2 presents the normalized contribution type distribution across the past 10 years for each filtering setting. Each stacked bar shows the year-normalized proportion of each contribution type. Across all settings, the most dominant types are D and B, with the exact balance differing by filter. In *all* and *MT*, type D has the largest share most years, while in *main*, type B is highly competitive—occasionally surpassing D. This indicates prevalence of engineering- and applications-oriented work in the field, which is unsurprising. Types G and C vary more significantly across settings: G is more prominent in *all* and *MT* than in *main*, whereas C appears more in *main*. This also fits with our expectations about the difference between main papers (for which acceptance is more selective) versus those in smaller conferences and workshops: the latter have a greater proportion of papers re-applying existing techniques to new data, whereas more technically substantial work that contributes mathematical and algorithmic innovations constitutes a larger proportion of main conference papers.

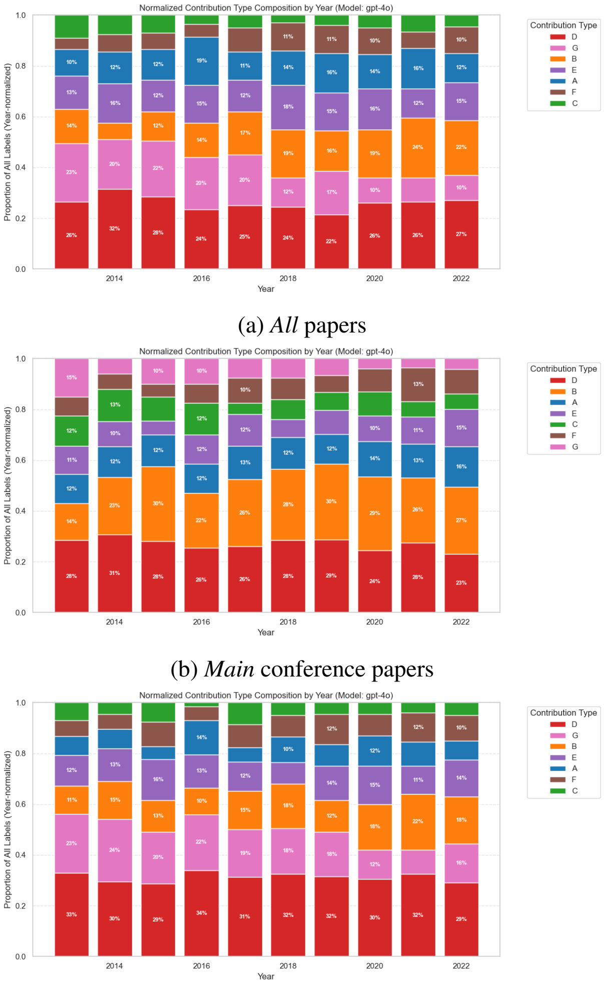
4.2 Temporal Trends in Contribution Types

Figure 3 illustrates the yearly number of papers assigned to each contribution type, with a consistent sampling of 100 papers per year per setting.

Across all settings, types D and B clearly dominate in terms of volume, though their trajectories diverge depending on the filtering criterion. In *MT*, type D remains the most frequent throughout the decade, whereas in *main*, type B exhibits a substantial rise after 2016—and remains high after 2019. This may correspond to increased focus on exploring different neural architectures for NLP, especially after the release of BERT (Devlin et al., 2019). Type F shows a marked increase after 2016 across all settings, perhaps reflecting growing importance of interpretability or evaluating popular models in different contexts or model. Type C remains consistently infrequent across all views, which makes sense - mathematical innovations are relatively rare in the field. Types A and E follow relatively stable, mid-range patterns—suggesting they maintain a persistent but modest role in the literature. Type G, once highly prevalent (especially in *MT*), declines steadily across all settings, suggesting a decreasing emphasis on this form of contribution. This is surprising - one would have imagined that work focusing on applications of existing models to new data would have risen in the age of data and benchmarking. However, there is perhaps a thin line between G and D with potential for label confusion, and it's possible that the decreasing G trend is evidence of another underlying trend or a model labeling artifact.

Limitations

The main limitation of this paper is that our study is conducted over a small sample set for a given year due to compute constraints, leading to perhaps unstable conclusions. These trends would



(a) *All* papers

(b) *Main* conference papers

(c) *MT*-focused papers

Figure 2: Normalized contribution type distribution by year for different filtering settings.

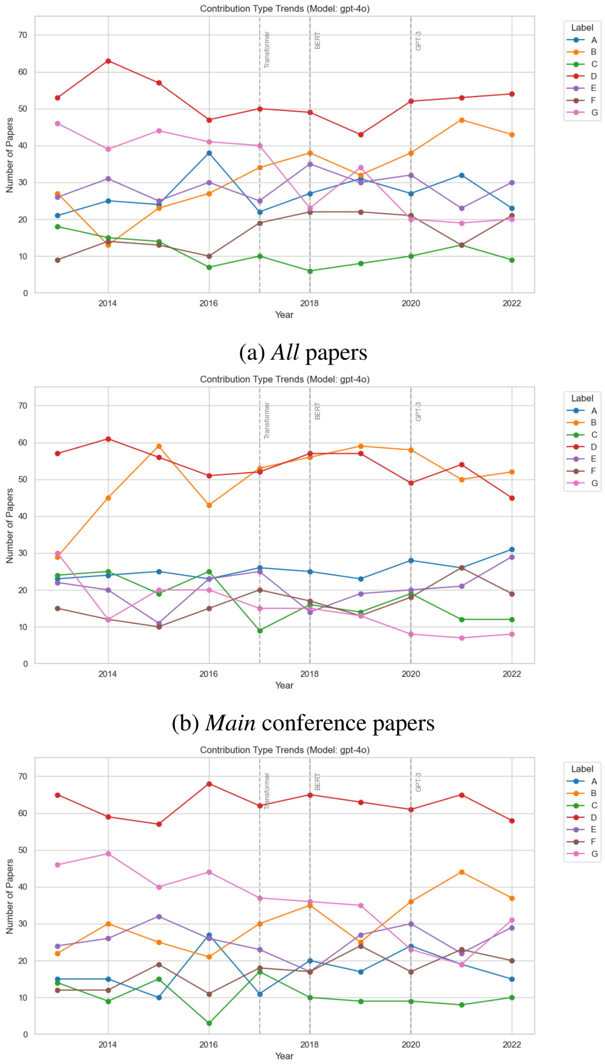
need to be validated over larger sample sets. Further, it would have also been nice to iterate on the annotation scheme with manual analysis of LLM outputs, with perhaps a second round of annotation post-refinement. We leave this to anyone who is interested in expanding this study :)

Contribution breakup

- Krirk: Code base implementation, data statistics, filtering, inference with LLMs, analysis
- Niyati: Project scoping, annotation scheme, analysis

Acknowledgments

We used OpenAI's ChatGPT to assist with grammar revision and to suggest relevant data analysis approaches. All final decisions and writing were done by the authors.



(a) *All* papers

(b) *Main* conference papers

(c) *MT*-focused papers

Figure 3: Temporal trends of contribution types for each filtering setting. Vertical lines mark key model releases.

A Data Fields

Column Name	Description
acl_id	Unique ACL ID
abstract	Abstract extracted by GROBID
full_text	Full text extracted by GROBID
corpus_paper_id	Semantic Scholar ID
pdf_hash	SHA1 hash of the PDF
numcitedby	Number of citations from Semantic Scholar
url	Link to publication
publisher	—
address	Address of conference
year	—
month	—
booktitle	—
author	List of authors
title	Title of paper
pages	—
doi	—
number	—
volume	—
journal	—
editor	—
isbn	—

Table 4: List of fields included in the ACL-OCL dataset. Dashes indicate fields with straightforward column names.

B Additional Dataset Statistics

We provide additional summary statistics of our dataset below. Figure 4 shows the number of papers per year, and Figure 5 shows the top 10 publication venues.

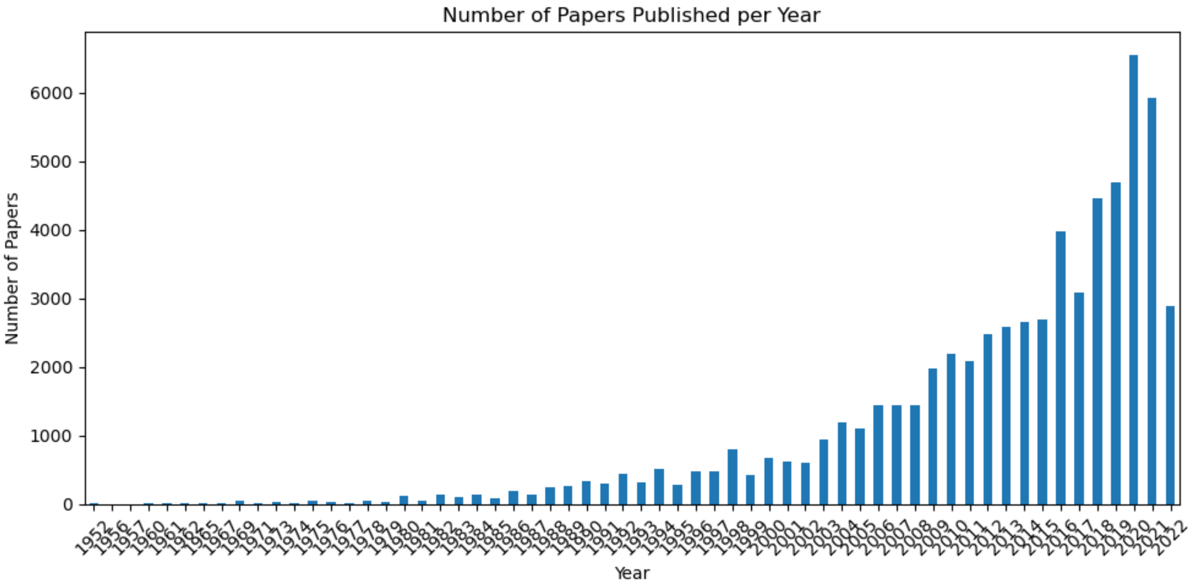


Figure 4: Number of papers per year in the filtered dataset.

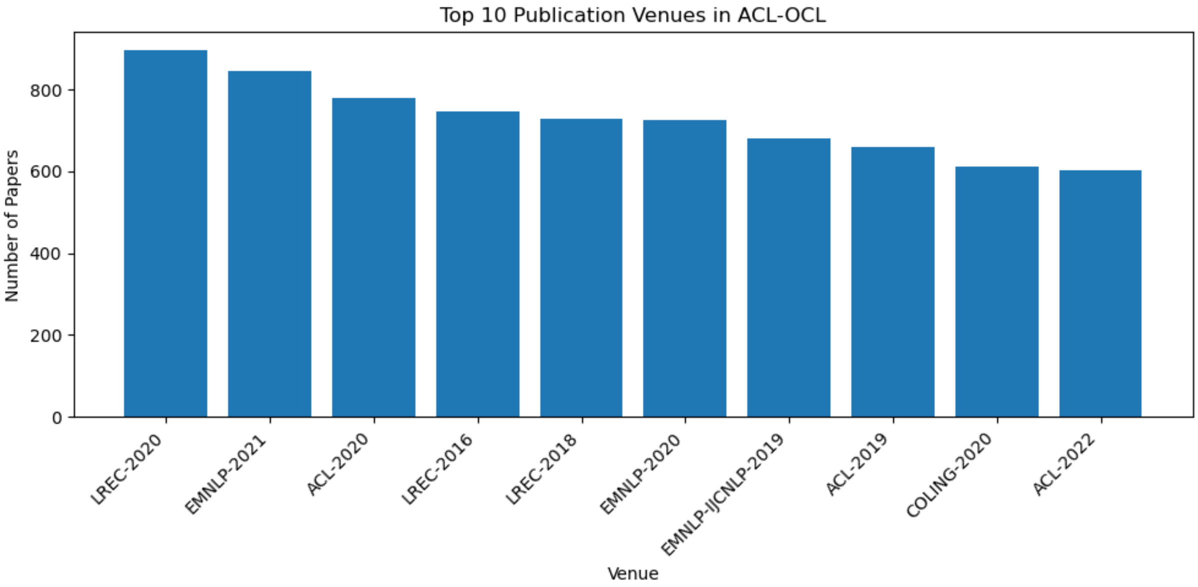


Figure 5: Top 10 publication venues by number of papers.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Antonios Dimakis, Stella Markantonatou, and Antonios Anastasopoulos. 2024. Dictionary-aided translation for handling multi-word expressions in low-resource languages. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2588–2595.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279.

Yi-Chung Lin and Keh-Yih Su. 2021. How fast can BERT learn simple natural language inference? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 626–633, Online. Association for Computational Linguistics.

Nelsi Melgarejo, Rodolfo Zevallos, Hector Gomez, and John E. Ortega. 2022. WordNet-QU: Development of a lexical database for Quechua varieties. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4429–4433, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Bo Zheng, Li Dong, Shaohan Huang, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Allocating large vocabulary capacity for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3203–3215.