Towards Multi-Modal Text-Image Retrieval to improve Human Reading

Florian Schneider and **Özge Alaçam** and **Xintong Wang** and **Chris Biemann**
Language Technology group
Universität Hamburg, Germany
florian.schneider-1@studium.uni-hamburg.de
{alacam|xwang|biemann}@informatik.uni-hamburg.de

Abstract

In primary school, children's books, as well as in modern language learning apps, multi-modal learning strategies like illustrations of terms and phrases are used to support reading comprehension. Also, several studies in educational psychology suggest that integrating cross-modal information will improve reading comprehension. We claim that state-of-the-art multi-modal transformers, which could be used in a language learner context to improve human reading, will perform poorly because of the short and relatively simple textual data those models are trained with. To prove our hypotheses, we collected a new multi-modal image-retrieval dataset based on data from Wikipedia. In an in-depth data analysis, we highlight the differences between our dataset and other popular datasets. Additionally, we evaluate several state-of-the-art multi-modal transformers on text-image retrieval on our dataset and analyze their meager results, which verify our claims.

1 Introduction

When we were babies, we learned our native language by combining our parents' words and visual hints. In primary school, children's books, as well as in modern language learning apps, like Babble[1] or Duolingo[2], this multi-modal learning strategy continues as illustrations of terms and phrases are used to support reading comprehension Also, multiple studies in educational psychology suggest that integrating cross-modal information will improve learning to read (Ecalle et al., 2009; Dalton and Grisham, 2011; Hahn et al., 2014; Gerbier et al., 2018; Kabooha and Elyas, 2018; Xie et al., 2019; Albahiri and Alhaj, 2020).

This paper presents initial research towards leveraging machine learning technology within a language learner context to improve human reading. In this scenario, the aim is to support a user's reading comprehension of arbitrary text by enhancing it with context-specific visual clues from state-of-the-art multi-modal transformers. The most popular training datasets for current models applied on text-image retrieval are MS COCO (Lin et al., 2014) and Flickr30k (Young et al., 2014; Plummer et al., 2015). Both datasets were created by crowdsourcing workers with the task to find short, simple and descriptive captions for images taken from Flickr[3]. We argue that sentences slightly advanced language learners might not comprehend are presumably more complex than the captions from COCO or Flickr30k. Hence we further claim that current models will perform poorly on more complex data.

The contributions of this work to verify these hypotheses are: *a*) the collection of a multi-modal dataset based on WikiCaps (Schamoni et al., 2018), which we call WISMIR (**WI**kiCaps **S**ubset for **M**ulti-Modal **I**mage-**R**etrieval); *b*) an in-depth analysis and comparison of WISMIR to other multi-modal datasets used for image-retrieval; *c*) a text-image retrieval evaluation of state-of-the-art image-retrieval models on WISMIR.

2 Related Work

During the last few years, there were significant breakthroughs in various computer vision tasks and models (Kirillov et al., 2020; Güler et al., 2018) as well as in the field of natural language processing. Especially with the recent dawn of transformers, models are increasingly capable of understanding texts semantics (Brown et al., 2020; Devlin et al., 2019; Yang et al., 2019). This progress of uni-modal models also led to a great leap forward in multi-modal visio-linguistic models, which are starting to leverage the power of transformers to work with text and images simultaneously. One

[1] https://babbel.com/
[2] https://duolingo.com/
[3] https://flickr.com/

of the several multi-modal tasks where these models pushed the boundaries is text-image retrieval, which we want to make use of in our language learner scenario. For this task, the model's learns a metric function $\Phi_{k,l} : \mathbb{R}^{|S_k| \times |I_l|} \rightarrow [0,1]$ that measures the similarity of sentence S_k and image I_l. The task's goal is to find the best matching image $I_k = \underset{l \in P}{\operatorname{argmax}} \Phi_{q,l}$ for a query sentence q from a pool of images P, the similarity scores $\{\Phi_{q,l} \mid l \in P\}$ have to be computed. The input to multi-modal transformers applied on text-image retrieval are textual tokens of a sentence S_k together with region features of an image I_l. Usually, textual tokens are generated by pre-trained BERT tokenizers (**?**). The visual region features are typically computed by pre-trained region and object detection and classification networks such as Faster-R-CNN with ResNet-101 (Ren et al., 2016; He et al., 2016; Anderson et al., 2018).

Multi-modal transformer networks can be grouped into so-called "early-fusion" models and "late-fusion" models. In early-fusion models such as UNITER (Chen et al., 2020), OSCAR (Li et al., 2020), or VL-BERT (Su et al., 2019), tokens of both modalities form the input to the network. Self-attention heads in the transformer-encoder layers (Vaswani et al., 2017) then compute joint-representations of both modalities, i.e., fine-grained word-region-alignments of the words $w \in S_k$ and the visual tokens $v \in I_l$. The similarity function $\Phi_{k,l}$ is an arbitrary combination of those joint-representations that depends on the respective model. Despite their remarkable performance of tasks on typical datasets like COCO or Flickr30k, early-fusion models are not applicable in real-world information retrieval systems with large pool of images because it would require tremendous computational power. As opposed to early-fusion models, in late-fusion models, the textual and visual modalities get forwarded through separate transformers for each modality. Later, the output of the textual transformer and the output of the visual transformer get fused depending on the model's specific implementation. For example, LXMERT (Tan and Bansal, 2019) and VilBERT (Lu et al., 2019) compute the fused cross-modality output with a third cross-modal transformer that takes the separate and uni-modal transformers' outputs as inputs. Other late-fusion models specially designed to solve multi-modal retrieval tasks like TERN (Messina et al., 2020) and TERAN (Nicola et al., 2020) use a more computationally efficient way.

3 Dataset Collection

The most popular datasets for pre-training and fine-tuning multi-modal transformers applied on retrieval tasks are MS COCO and Flickr30k. Both are hand-crafted datasets, with short, descriptive and conceptual captions created by crowdsourcing workers describing mostly non-iconic images from Flickr. Within a language learner scenario, we argue that the sentences a user does not understand while reading are presumably more complex than the short and relatively simple caption sentences from COCO or Flickr30k. Consequently, we claim that models trained with this data will perform worse on more complex textual data.

An example of a multi-modal dataset containing non-constrained and heterogeneous text-image pairs is WikiCaps (Schamoni et al., 2018), which contains about 3.8 million images and their respective English captions from Wikipedia articles. The authors of WikiCaps only provide a tab-separated file containing the Wikimedia file IDs of the image and the respective caption together with a perl script to download the images serially. To make the data more accessible, we developed an efficient python application, which we released on GitHub[4]. This tool is capable of collecting corpus statistics based on the captions using different models and frameworks, flexibly filtering the data with user-defined filters, downloading the images in parallel, applying customizable transformations to the images, and finally, persisting the data in an easy to use and efficient format. Using the tool, we collected and released[5] a first proof-of-concept dataset, which we call we call WISMIR (**WI**kiCaps **S**ubset for **M**ulti-Modal **I**mage-**R**etrieval), containing 187598 text-image pairs. We randomly split the data in a training and a test set containing $178218 (95\%)$ and $9380 (5\%)$, respectively.

3.1 Data Analysis

To examine the differences of COCO, Flickr30k, and WISMIR, we used our tool to generate the corpus statistics discussed in the following. Since the models used to generate this data are not flawless, we utilized three different frameworks[6], namely

[4] `https://git.io/Jtw2P`

[5] `https://git.io/JscGV`

[6] `https://spacy.io/`; `https://www.nltk.org/`; `http://polyglot-nlp.com/`

spaCy, NLTK, and Polyglot, to get a more reliable impression on the distribution of the data. Please note that we only show the two most notable distinctions of WISMIR, COCO, and Flickr30k in the following due to this paper's brevity. [7]

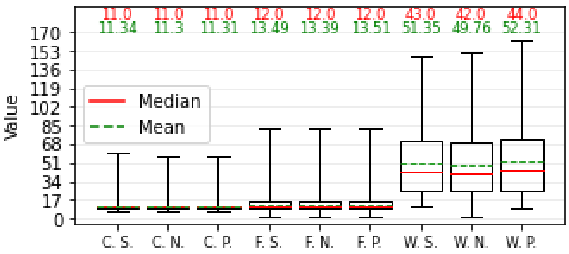

Figure 1: Boxplot diagrams[7] for the **number of tokens per caption** in COCO, Flickr30k, and WISMIR, generated by different tokenization models.

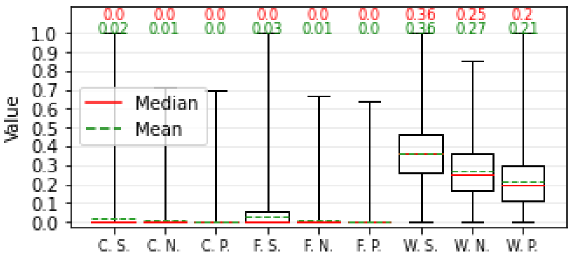

Figure 2: Boxplot diagrams[7] for the **ratio of tokens contained in named entities and all tokens of a caption** in COCO, Flickr30k, and WISMIR, generated by different tokenizers and named entity recognition models.

Figure 1 shows that the average number of tokens per caption is between $3.6 - 4.6$ times higher in WISMIR than in COCO or Flickr30k. This is an essential property because language learners will most probably have more difficulties comprehending long paragraphs than short paragraphs. The most significant difference between the datasets is shown in Figure 2: In COCO and Flickr30k, there are almost no named entities, while in WISMIR, between $21 - 36$ % of a captions' tokens are related to named entities on average.

3.1.1 Readability Comparison

To further underline the differences between COCO, Flickr30k, and WISMIR, we computed the Flesch-Kincaid (Farr et al., 1951) (FK) and Dale-Chall (Chall and Dale, 1995) (DC) readability scores for random samples of the datasets containing $10^6 \pm 0.1\%$ characters. Because these readability scores depend on the number of sentences, words, and syllables in the text, counted by imperfect models, we use two different implementations[8] to obtain more reliable results. In Figure 3, we can observe that the captions of COCO and Flickr30k should be easily understood by an average 4th to 6th-grade US student. In contrast, WISMIR captions are recommended for college students or higher, according to the FK and DC scores.

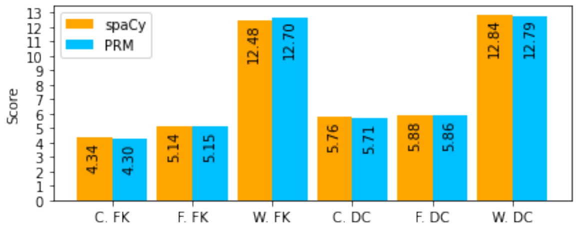

Figure 3: Comparison of Flesch-Kincaid (FK) and Dale-Chall (DC) readability scores of randomly sampled subsets of COCO (C), Flickr30k (F), and WISMIR (W) captions containing $10^6 \pm 0.1\%$ characters computed by two different frameworks.

4 Model Evaluations

To verify our claim that models pre-trained solely on COCO and Flickr30k perform poorly on text-image retrieval with more complex and heterogeneous data like WISMIR, we conducted several evaluations.

As listed in Table 1, evaluation scores on WISMIR, especially for COCO and Flickr30k pre-trained TERAN models, are meager and fall way below our expectations. It seems to be the case that COCO and Flickr30k did not contribute anything meaningful in the models' training process when using them for text-image retrieval on WISMIR. The same appears to be true for the other way around, i.e., TERAN models trained on WISMIR perform very badly on COCO and Flickr30k. Even $\text{UNITER}_{\text{base}}$ pre-trained with much more data (5.6M samples) from COCO, Visual Genome (Krishna et al., 2017), Conceptual Captions (Sharma et al., 2018), and SBU Captions (Ordonez et al., 2011) performed poorly. While the results are better for TERAN_{W}, they are still far behind the re-

[7] On the x-axis, the name of the dataset and the framework used to generate the statistics the respective boxplot stands for are abbreviated by two letters: C for COCO, F for Flickr30k, W for WISMIR, S for spaCy, N for NLTK, and P for polyglot.

[8] https://git.io/JtgjK; https://git.io/JtgPG)

Model	Data	R@1	R@5	R@10
TERAN_W	W	**8.9**	**26.9**	**38.2**
TERAN_C	W	1.1	3.7	5.6
TERAN_F	W	0.9	2.7	4.4
$\text{UNITER}_\text{base}$	W	5.31	13.28	18.75
TERAN_W	C	2.0	6.9	11.5
TERAN_C	C	42.6	72.5	82.9
$\text{UNITER}_\text{base}$	C	**50.33**	**78.52**	**87.16**
TERAN_W	F	4.6	14.5	22.6
TERAN_F	F	59.4	84.8	90.5
$\text{UNITER}_\text{base}$	F	**72.52**	**92.36**	**96.08**

Table 1: Recall@K evaluation results of different models on text-image retrieval on multiple test sets. The letters C for COCO, F for Flickr30k, or W for WISMIR are the datasets' abbreviations. In the subscripts, it indicates the training data of the TERAN model.

sults of TERAN_C or TERAN_F and UNITER on COCO or Flickr30k.

4.1 Error Analysis

To ensure that the poor performance of $TERAN_W$ does not originate from an eventual imbalance between the train and test set, we compared the data distribution between different subsets of WISMIR and found that the differences are neglectable.

Further, we found that the model has seen that 84% of the token types, 72% of the noun token types, and 80% of the named entity types of the test set during training. From these findings, we can conclude that the difficulties with WISMIR do not originate from surface forms of the dataset's captions but from a deeper semantic or discourse level.

Further problems could be introduced by the large number of tokens per caption on average. Most of the words in a lengthy caption are probably not grounded in an image region and can therefore be regarded as noise for word-region-alignments. When too many words are not depictable or are not grounded in ROIs, it leads to loose coupling between the caption and the image, which is clearly not beneficial for the models' training.

Other sources of issues might lie in the architecture or the training process of the models. All weights of the model are trained via hinge-based triplet-loss leveraging global image-caption similarity scores computed by pooling matrices that contain the cosine-similarities between the visual and textual contextual embeddings. For long sentences with many words and a limited number of 36 visual tokens per image, it could be challenging to sample good (anchor, positive, negative) triplets required by the loss function and finally cause problems while training the model.

4.2 Future Experiments

As described in the previous sections, we identified multiple obstacles that need to be overcome to leverage multi-model transformers like TERAN for real-world information retrieval systems within a language learner context. Several experiments are planned for future work to tackle the issues: We will collect a new version of WISMIR where we augment the named entities in the data with their corresponding labels ("PER", "ORG", etc.), and further increase the size of the dataset to examine the number of samples at which the performance does no longer improve. We will train and evaluate a new TERAN model on the improved WISMIR version to verify that the performance improves.

Text-image retrieval is hard to evaluate because the quality of the models' outcomes is subjective, and there are multiple relevant and "correct" images for a given query. To overcome this issue, non-exact metrics like DCG or NDCG, which rely on relevance scores between the model results, are often used to evaluate information retrieval systems. The problem we are faced with is that there is no straightforward solution to compute these relevance scores for WISMIR. Therefore we plan a small-scale user study on Amazons' crowdsourcing platform, MTurk[9], to let humans assess the models' performances.

5 Conclusion

In this paper, we verify our claim that multi-modal transformers for text-image retrieval, pre-trained on common datasets like COCO and Flickr30k, cannot generalize well on more complex textual data. Therefore, we collected a multi-modal image-retrieval dataset, WISMIR, and conducted several analysis experiments that underline its differences to COCO and Flickr30k. Additionally, we evaluated two state-of-the-art multi-modal transformers on text-image retrieval on this novel dataset to verify our claim. We discovered significant problems the evaluated models have with our dataset and in the dataset itself, which we will approach in future work.

[9] https://www.mturk.com/

Acknowledgements

This research was partially funded by the German Research Foundation – DFG Transregio SFB 169: Cross-Modal Learning

References

Mohammed H. Albahiri and Ali A. M. Alhaj. 2020. Role of visual element in spoken English discourse: implications for YouTube technology in EFL classrooms. *The Electronic Library*, 38(3):531–544.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, Salt Lake City, UT, USA.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, Virtual.

Jeanne S. Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120, Virtual.

Bridget Dalton and Dana L. Grisham. 2011. eVoc Strategies: 10 Ways to Use Technology to Build Vocabulary. *Reading Teacher*, 64(5):306–317.

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, Minneapolis, MN, USA.

Jean Ecalle, Annie Magnan, Houria Bouchafa, and Jean Emile Gombert. 2009. Computer-Based Training with Ortho-Phonological Units in Dyslexic Children: New Investigations. *Dyslexia*, 15(3):218–238.

James N. Farr, James J. Jenkins, and Donald G. Paterson. 1951. Simplification of Flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Emilie Gerbier, Gérard Bailly, and Marie L. Bosse. 2018. Audio–visual synchronization in reading while listening to texts: Effects on visual behavior and verbal learning. *Computer Speech & Language*, 47:74–92.

Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, Salt Lake City, UT, USA.

Noemi Hahn, John J. Foxe, and Sophie Molholm. 2014. Impairments of multisensory integration and cross-sensory learning as pathways to dyslexia. *Neuroscience & Biobehavioral Reviews*, 47:384–392.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, Las Vegas, NV, USA.

Raniah Kabooha and Tariq Elyas. 2018. The Effects of YouTube in Multimedia Instruction for Vocabulary Learning: Perceptions of EFL Students and Teachers. *English Language Teaching*, 11(2):72–81.

Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. 2020. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, Virtual.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137, Virtual.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755, Zurich, Switzerland.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *Advances in Neural Information Processing Systems*, volume 32, pages 13–23, Vancouver, Canada.

Nicola Messina, Fabrizio Falchi, Andrea Esuli, and Giuseppe Amato. 2020. Transformer Reasoning Network for Image-Text Matching and Retrieval. *arXiv preprint arXiv:2004.09144*.

Messina Nicola, Amato Giuseppe, Esuli Andrea, Falchi Fabrizio, Gennaro Claudio, and Marchand-Maillet Stéphane. 2020. Fine-grained Visual Textual Alignment for Cross-Modal Retrieval using Transformer Encoders. *arXiv preprint arXiv:2008.05231*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2Text: Describing Images Using 1 Million Captioned Photographs. In *Advances in Neural Information Processing Systems*, volume 24, pages 1143–1151, Granada, Spain.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, Santiago, Chile.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149.

Shigehiko Schamoni, Julian Hitschler, and Stefan Riezler. 2018. A Dataset and Reranking Method for Multimodal MT of User-Generated Image Captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 140–153, Boston, MA, USA.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. VL-BERT: Pretraining of Generic Visual-Linguistic Representations. *arXiv e-prints*, pages arXiv–1908.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. 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 5103–5114, Hong Kong, China.

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*, 30:5998–6008.

Heping Xie, Richard E. Mayer, Fuxing Wang, and Zongkui Zhou. 2019. Coordinating Visual and Auditory Cueing in Multimedia Learning. *Journal of Educational Psychology*, 111(2):235–255.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.