Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4285–4290
Marseille, 20-25 June 2022

CEPOC: The Cambridge Exams Publishing Open Cloze dataset

Mariano Felice, Shiva Taslimipoor, Øistein E. Andersen, Paula Buttery
ALTA Institute, Computer Laboratory, University of Cambridge
Cambridge, UK
{mf501, st797, oa223, pjb48}@cam.ac.uk

Abstract

Open cloze tests are a standard type of exercise where examinees must complete a text by filling in the gaps without any given options to choose from. This paper presents the Cambridge Exams Publishing Open Cloze (CEPOC) dataset, a collection of open cloze tests from world-renowned English language proficiency examinations. The tests in CEPOC have been expertly designed and validated using standard principles in language research and assessment. They are prepared for language learners at different proficiency levels and hence classified into different CEFR levels (A2, B1, B2, C1, C2). This resource can be a valuable testbed for various NLP tasks. We perform a complete set of experiments on three tasks: gap filling, gap prediction, and CEFR text classification. We implement transformer-based systems based on pre-trained language models to model each task and use our dataset as a test set, providing promising benchmark results.

Keywords: open cloze, blank-filling, language learning, second language testing, Cambridge examinations

1. Introduction

The cloze test (Taylor, 1953) is a standard testing procedure where certain words are replaced with gaps in a piece of text, which must then be filled by the student. There are different variations of the original cloze test:

Open cloze No options are given to fill in the blank. Students must produce the answer with no help.

Multiple choice An answer must be chosen from a set of given options containing the *key* (i.e. correct answer) and *distractors*.

C-test A modified version of the open cloze test where the first few letters of the answer are kept in the gap as a hint (Raatz and Klein-Braley, 1981).

An example of each cloze type is given in Figure 1.

Cloze tests are widely used for testing language proficiency due to their simplicity and efficiency (Fotos, 1991; Jonz, 1991; Tremblay, 2011; Trace, 2020). However, they must be carefully designed in order to provide accurate measures of ability, often requiring calibration using methods such as those from Item Response Theory (IRT) (Lord, 1980).

Although there is vast research into modelling multiple-choice cloze tests (mainly in the field of reading comprehension, e.g. Xie et al. (2018) or Kurdi et al. (2020)), to the best of our knowledge, public datasets of calibrated cloze tests are non-existent, hindering research into automated testing. In this paper, we attempt to address this problem by making the following contributions: 1) we release the first dataset of expertly designed and calibrated open cloze tests at different proficiency levels in English, 2) we provide benchmark results for three tasks: gap generation, gap filling and full-text proficiency level prediction, and 3) we offer insights into how our dataset can help further research into language learning and assessment.

The rest of this paper is organised as follows: Section 2 briefly discusses related work on the available cloze datasets. In Section 3, we detail the composition of our CEPOC dataset, its source, statistics and some lexical analysis. Section 4 elaborates on our experiments on three interesting applications of the dataset. We describe how we model the tasks and report the results of state-of-the-art transformer-based models that can be used as promising benchmarks for the three tasks. Section 5 describes other possible applications of CEPOC and Section 6 concludes with a summary of our work.

(i) Genealogy is a branch .... history.

(ii) I have a degree .... mechanical engineering.
a) on b) in c) for d) about

(iii) China is the largest pro........ of garlic.

Figure 1: An example of (i) open cloze, (ii) multiple choice and (iii) c-test.

2. Related Work

Most work on cloze tests is aimed at reading comprehension or question answering, with less emphasis on second language learning. The Microsoft Research Sentence Completion Challenge, for example, consists of 1,040 multiple choice sentences from Sherlock Holmes stories where a content word is blanked and four alternatives are given (Zweig and Burges, 2012). The CNN/Daily Mail Reading Comprehension Task (Hermann et al., 2015), on the other hand, presents short newspaper passages with a summary that must be filled with an entity from the text. The task served as inspiration for other similar datasets such as the “Who-did-What” dataset (Onishi et al., 2016), the Children’s Book Test (Hill et al., 2016) for the comprehension of children’s stories, the People Daily and Children’s Fairy Tale dataset (Cui et al., 2016) for Chinese, and BioRead (Pappas et al., 2018) and BIOMRC (Pappas et al., 2020) for the biomedical domain.

Exam	CEFR level	# tests	Vocab. size	Avg. # tokens	Avg. TTR
Key (KET)	A2	6	290	128	65.21%
Preliminary (PET)	B1	21	854	163	61.91%
First (FCE)	B2	36	1,759	176	61.84%
Advanced (CAE)	C1	30	1,562	191	62.23%
Proficiency (CPE)	C2	21	1,270	192	63.56%

Table 1: Dataset composition. The average number of tokens and Type-Token Ratio (TTR) are per task.

# gaps	KET	PET	FCE	CAE	CPE
6	-	9	-	-	-
8	-	12	-	-	-
9	-	-	9	8	15
10	-	-	18	7	1
11	4	-	9	8	5
12	2	-	-	6	-
13	-	-	-	1	-
Total tasks	6	21	36	30	21

Table 2: Distribution of tasks by number of gaps.

Cloze question datasets for language learners are less common. The Chengyu Cloze Test (Jiang et al., 2018) and ChID (Zheng et al., 2019) datasets contain short passages and sentences where a Chinese idiom has been removed and has to be chosen from a list. For English, the CLOTH dataset (Xie et al., 2018) is probably the closest to our proposal, since each exercise contains a full-text passage with multiple gaps. However, gaps in CLOTH are multiple choice questions that test a range of abilities such as reasoning, grammar and paraphrasing. The HyTeC-cloze system (Kleijn et al., 2019) also creates multiple gaps for full-text passages but, unlike CLOTH, they are open cloze items and only available for Dutch.

Unlike previous work, the CEPOC dataset we present in this paper comprises full-passage open cloze tests containing multiple gaps each. These gaps are specifically aimed at testing English grammar and vocabulary, have been calibrated using IRT and are graded by proficiency level.

3. CEPOC Dataset Composition

3.1. The Open Cloze Test

Whereas existing datasets focus on multiple choice tests, the Cambridge Exams Publishing Open Cloze (CEPOC) dataset is novel in providing *open* cloze tests and across a range of proficiency levels. CEPOC thus constitutes a more challenging dataset since answers are not given in advance as a list of candidate options and must be produced from scratch. This not only enables the testing of productive skills in students (Pino et al., 2008) but also serves as a more realistic benchmark for testing language model prediction capabilities (Donahue et al., 2020; Gonçalo Oliveira, 2021).

# answers	KET	PET	FCE	CAE	CPE
1	54	126	296	241	159
2	12	18	45	37	31
3	2	5	16	6	4
4	-	1	2	3	5
5	-	-	1	1	1
Total gaps	68	150	360	288	200

Table 3: Distribution of gaps per exam with a specific number of valid answers.

A bicycle you can fold up

Folding bicycles have *been* around for quite some time now. However, an amazing new Japanese version ____ be folded with a swiftness and efficiency never seen before. This bike is designed ____ that it is possible to fold it up quickly. Once folded, you pull the bike along ____ ease.

Figure 2: Shortened sample from CEPOC (FCE level).

CEPOC contains a collection of open cloze tasks that have been previously published as practice material by Cambridge Exams Publishing, a division of Cambridge University Press & Assessment (CUP&A)[1]. CUP&A's English examinations span five different proficiency levels of the Common European Framework of Reference (CEFR) for languages (Council of Europe, 2001), ranging from A2 (elementary) to C2 (proficient). Each examination tests four basic skills (reading, writing, listening and speaking) using a variety of tasks (word completion, direct questions, multiple matching, multiple choice, open cloze tests, etc.). These tasks are created by expert 'item writers' following strict design principles (ALTE, 2005; ALTE, 2011) and are further validated and calibrated using standard IRT procedures (Corrigan and Crump, 2015). CEPOC only focuses on open cloze tasks from the 'Reading' and 'Use of English' sections of the exams.

3.2. Dataset Statistics

Each open cloze task in CEPOC consists of a short written passage that is adapted to the CEFR level of the corresponding examination. Details of CEPOC's composition are reported in Table 1.

Each test contains a variable number of blanks or 'gaps' depending on the level, as shown in Table 2. Each of these gaps must be filled by a single word and may allow more than one possible answer, with a maximum of 5. The distribution of possible answers per gap is shown in Table 3. The first gap in each test is an example which is included for illustrative purposes only and should not be used for testing since it has not been calibrated. Figure 2 shows a shortened sample test.

3.3. Gap analysis

CEPOC comprises tests for a range of CEFR proficiency levels, enabling cross-level comparisons. Firstly, we report gap distribution statistics in Table 4. The distance between gaps is fairly homogeneous at around 20 tokens, except for KET were it is roughly

[1] `https://www.cambridge.org/`

	KET	PET	FCE	CAE	CPE
Avg. gap distance	10.81	19.36	17.74	19.23	19.49
Avg. gaps/sentence	0.97	0.69	1.15	1.07	1.14
Gaps at the beginning	27	57	116	111	66
Gaps in the middle	26	48	119	85	68
Gaps at the end	15	45	125	92	66

Table 4: Gap distribution and location in the sentences.

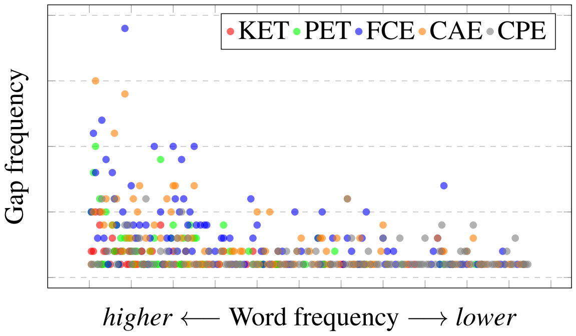

Figure 3: Frequency of words in the English language (*x* axis, decreasing) vs. their frequency as gaps in the different exams represented in CEPOC (*y* axis).

10, likely due to the smaller size of the passages. The average number of gaps per sentence is roughly 1 in all cases however. When splitting each sentence into three equal-sized parts, we observe that gaps are fairly equally distributed between the beginning, middle and end of the sentence. However, over 60% of gaps tend to occur from the middle onwards, suggesting that they require more context. By design, no gaps are created for the first token in the passage.

Secondly, we carry out an analysis of the gapped words and their frequency by level. Figure 3 shows that gaps in examinations at the lower CEFR levels tend to be very high frequency words in the English language[2] (*a*, *to*, *the*, *if*), while gaps at higher levels feature less frequent vocabulary (*whose*, *regardless*, *extent*). This is in line with the expectations that students with higher proficiency should be able to use a wider vocabulary.

Lastly, we look into the parts of speech (PoS) of the words that are gapped at each CEFR level. The distribution of PoS tags is shown in Table 5.[3]

In general, the majority of gaps correspond to function words (prepositions, determiners, conjunctions, etc.) although this proportion tends to decrease moderately at higher CEFR levels in favour of content words (adverbs, adjectives and nouns). This behaviour is by design: while elementary students are expected to demonstrate a good use of grammar, more advanced students are expected to know a wider range of lexico-grammatical items, such as idioms and collocations.

Gap PoS	KET	PET	FCE	CAE	CPE
ADJ	2	4	16	6	17
ADP	13	33	74	86	42
ADV	6	13	51	41	38
AUX	8	14	39	23	4
CCONJ	2	2	9	11	6
DET	10	23	50	34	22
INTJ	0	0	1	1	0
NOUN	0	6	12	5	16
NUM	0	3	10	4	2
PART	5	10	6	8	2
PRON	10	18	34	19	8
SCONJ	8	21	50	34	18
VERB	4	3	8	16	25
Function words	88%	85%	78%	82%	64%
Content words	12%	15%	22%	18%	36%

Table 5: PoS distribution of the gaps in CEPOC.

Model	KET	PET	FCE	CAE	CPE
bert-base-uncased	95.45	97.99	95.51	89.58	85.20
bert-large-uncased	94.03	98.00	97.77	92.01	90.36
roberta-base	98.53	99.33	96.38	94.79	93.50
roberta-large	100.00	99.33	98.33	97.92	96.50

Table 6: Accuracy of different masked language models on the “gap-filling” task.

4. Experiments

We show three possible applications of our corpus by using it as a benchmark dataset for three tasks: 1) gap filling, 2) gap prediction and 3) text-level CEFR level estimation. Each task is modelled using fine-tuned transformer-based systems and evaluated on CEPOC, as described in the following sections. While we encourage the use of CEPOC as a benchmark, it also lends itself to other uses such as fine-tuning or corpus analysis.

4.1. Gap Filling

We use two state-of-the-art (SOTA) pre-trained language models, BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), for the task of solving our open cloze tests automatically.[4] These masked language models are naturally well-suited to the “gap-filling” task, so we apply them to our dataset in order to provide a benchmark. We use the top prediction from each model as the predicted answer.

Results in Table 6 show that both BERT and RoBERTa can solve our open cloze tasks with exceptional accuracy, reaching over 90% in the majority of cases. As expected, performance decreases at the higher levels, confirming that test difficulty increases as we go up the CEFR scale. We also observe that larger models are able to answer more gaps thanks to their bigger vocabulary size, with RoBERTa consistently outperforming BERT, especially at the higher levels.

Prediction errors can occur for gaps with modal verbs (*the lighting* {*must/should* ✓ *can* ✗} *be right*), multiple answers ({*With/In/After* ✓ *During* ✗} *the aerobic exer-*

[2] Estimated using the wordfreq Python library: `https://github.com/rspeer/wordfreq`.

[3] PoS tagging was done using spaCy v2 (`https://spacy.io`), which uses the Universal Dependencies tagset: `https://universaldependencies.org/u/pos/`.

[4] From Hugging Face: `https://huggingface.co`.

Exams	KET	PET	FCE	CAE	CPE
# gaps	68	150	360	288	200
Random baseline	23.28	12.58	14.95	14.79	11.00
Exercise Maker	-	-	23.17	15.28	10.30
ELECTRA model	60.29	40.67	53.89	47.22	50.00

Table 7: Accuracy of gap prediction systems on the different sections of the CEPOC dataset.

cise), long-range dependencies (*But that hasn't stopped the self-attaching Post-it note ...* {*from* ✓*and* ✗} *becoming an essential piece...*) or where the model prioritises a more frequent phrase (*at* {*worst* ✓*times* ✗}, {*without* ✓*in* ✗} *question*). Though less frequently, RoBERTa also struggles with gaps BERT has difficulty with.

4.2. Gap Prediction

We also built a system that attempts to automatically create open cloze tests from a text passage by predicting which words would make good potential gaps. The task is modelled as a supervised sequence tagging problem where each token is classified as being a good potential gap or not. We employ ELECTRA's discriminator (Clark et al., 2020), a SOTA pre-trained language model. ELECTRA's objective for pre-training is to detect the tokens that are randomly replaced by a generator, rather than generating words for masked tokens (as in BERT).[5] We use ELECTRA's discriminator by adding a linear layer on top for token classification. A model is built for each exam in CEPOC and trained on a larger private collection of open cloze tests. This training data contains 267 tasks for KET, 180 for PET, 356 for FCE, 281 for CAE and 146 for CPE.

Our model is compared to two other systems: a) a random baseline, that randomly predicts gaps based on PoS gap frequency for each exam, and b) Exercise Maker (Malafeev, 2014), a rule-based system based on the most frequently gapped words in CUP&A's exams. All systems are set to predict the same number of gaps per task as they have in the gold standard. We evaluate our prediction results based on a strict matching between the gaps predicted by our models and those in the gold standard. Table 7 reports accuracy for all our systems, fine-tuned separately for each exam. Results for the random baseline and Exercise Maker correspond to the average of 5 runs. Exercise Maker's generation mode was set to the appropriate exam in each case but was not available for KET nor PET.

As expected, results show that the random baseline has very low performance and is only slightly outperformed by Exercise Maker on FCE and CAE. Performance of both systems, however, decreases as the CEFR level goes up, suggesting that gaps are less predictable at higher proficiency levels. Our ELECTRA model outperforms the other systems by a large margin and, while results are less homogeneous, they also exhibit decreasing accuracy. Further experiments involving human evaluation show that many gaps predicted by our model are also equally useful, despite not being matched by the gold standard (Felice et al., 2022).

4.3. CEFR Classification

For this experiment, we built a classification model to classify text passages into different CEFR proficiency levels (A2, B1, B2, C1 and C2). We employ the SOTA pre-trained transformer-based sequence classification model RoBERTa and fine-tune it for a few epochs using the same training data as in Section 4.2. We feed the text passages of our open cloze tests, where gaps are filled with their first acceptable answer, as input to the classifier and label the texts with the CEFR level of the exam they were extracted from.

The accuracy of automatically assigning all 114 tasks in CEPOC to their CEFR level (as described in Table 1) is 94.55% for our model vs. 20% for random and 32% for the majority class. This shows that tests at each level are clearly distinguishable from the rest, confirming that they have been carefully tailored to the needs of students at each level.

5. Research Directions

CEPOC fills a gap in the realm of automated second language learning, where public datasets of calibrated proficiency-graded open cloze tests are non-existent. As mentioned in Section 4, we encourage the use of CEPOC as a benchmark dataset but believe it can be exploited in other ways too. Some potential applications we envisage include:

- using the tests in CEPOC with cohorts of students to collect gap difficulty statistics, which could be used to perform gap difficulty prediction;
- characterising students' expected proficiency at each CEFR level by analysing the grammatical structures, lexical complexity and other linguistic indicators in the tests;
- investigating the factors that determine the effectiveness of gaps, such as context, distance to other gaps, entropy (Felice and Buttery, 2019), etc.

CEPOC is free to use for research purposes and is available at `https://github.com/CambridgeALTA/cepoc`.

6. Conclusion

This paper presented CEPOC, the first dataset of open cloze tests for learners of English at different CEFR levels. The tests in CEPOC have been designed and calibrated following strict procedures and are part of preparation materials for well-known English proficiency examinations.

We described how CEPOC could serve as an ideal dataset for a number of applications and provided encouraging benchmark results for three tasks: gap filling, gap prediction, and CEFR text classification. CEPOC is free to use for research purposes.

[5] In our experiments on FCE, ELECTRA shows better performance than RoBERTa for token classification.

7. Bibliographical References

Association of Language Testers in Europe (ALTE). (2005). Materials for the Guidance of Test Item Writers. Technical report, Association of Language Testers in Europe, July.

Association of Language Testers in Europe (ALTE). (2011). Manual for Language Test Development and Examining. Technical report, Association of Language Testers in Europe.

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

Corrigan, M. and Crump, P. (2015). Item Analysis. *Research Notes 59*, pages 4–9.

Council of Europe. (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, Cambridge.

Cui, Y., Liu, T., Chen, Z., Wang, S., and Hu, G. (2016). Consensus Attention-based Neural Networks for Chinese Reading Comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, Minnesota, June. Association for Computational Linguistics.

Donahue, C., Lee, M., and Liang, P. (2020). Enabling Language Models to Fill in the Blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online, July. Association for Computational Linguistics.

Felice, M. and Buttery, P. (2019). Entropy as a Proxy for Gap Complexity in Open Cloze Tests. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 323–327, Varna, Bulgaria, September. INCOMA Ltd.

Felice, M., Taslimipoor, S., and Buttery, P. (2022). Constructing Open Cloze Tests Using Generation and Discrimination Capabilities of Transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, May.

Fotos, S. S. (1991). The Cloze Test as an Integrative Measure of EFL Proficiency: A Substitute for Essays on College Entrance Examinations?*. *Language Learning*, 41(3):313–336.

Gonçalo Oliveira, H. (2021). ”Answering Fill-in-the-Blank Questions in Portuguese with Transformer Language Models”. In Goreti Marreiros, et al., editors, *Progress in Artificial Intelligence*, pages 739–751, Cham. Springer International Publishing.

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching Machines to Read and Comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS’15, page 1693–1701, Cambridge, MA, USA. MIT Press.

Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations. In Yoshua Bengio et al., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Jiang, Z., Zhang, B., Huang, L., and Ji, H. (2018). Chengyu Cloze Test. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 154–158, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jonz, J. (1991). Cloze item types and second language comprehension. *Language Testing*, 8(1):1–22.

Kleijn, S., Maat, H. P., and Sanders, T. (2019). Cloze testing for comprehension assessment: The HyTeC-cloze. *Language Testing*, 36(4):553–572.

Kurdi, G., Leo, J., Parsia, B., Sattler, U., and Al-Emari, S. (2020). A Systematic Review of Automatic Question Generation for Educational Purposes. *I. J. Artificial Intelligence in Education*, 30(1):121–204.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Malafeev, A. (2014). Language Exercise Generation: Emulating Cambridge Open Cloze. *Int. J. Concept. Struct. Smart Appl.*, 2(2):20–35, July.

Onishi, T., Wang, H., Bansal, M., Gimpel, K., and McAllester, D. (2016). Who did What: A Large-Scale Person-Centered Cloze Dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas, November. Association for Computational Linguistics.

Pappas, D., Androutsopoulos, I., and Papageorgiou, H. (2018). BioRead: A New Dataset for Biomedical Reading Comprehension. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Pappas, D., Stavropoulos, P., Androutsopoulos, I., and McDonald, R. (2020). BioMRC: A Dataset for

Biomedical Machine Reading Comprehension. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 140–149, Online, July. Association for Computational Linguistics.

Pino, J., Heilman, M., and Eskenazi, M. (2008). A Selection Strategy to Improve Cloze Question Quality. *Intelligent Tutoring Systems for Ill-Defined Domains: Assessment and Feedback in Ill-Defined Domains.*, page 22.

Raatz, U. and Klein-Braley, C. (1981). The C-Test–A Modification of the Cloze Procedure. In Christine Klein-Braley Terry Culhane et al., editors, *Practice and Problems in Language Testing 7*, pages 113–138. University of Essex, Dept. of Language & Linguistics, Colchester, England.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Trace, J. (2020). Clozing the gap: How far do cloze items measure? *Language Testing*, 37(2):235–253.

Tremblay, A. (2011). Proficiency assessment standards in second language acquisition research:"Clozing" the gap. *Studies in Second Language Acquisition*, 33(3):339–372.

Xie, Q., Lai, G., Dai, Z., and Hovy, E. (2018). Large-scale Cloze Test Dataset Created by Teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium, October-November. Association for Computational Linguistics.

Zheng, C., Huang, M., and Sun, A. (2019). ChID: A Large-scale Chinese IDiom Dataset for Cloze Test. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 778–787, Florence, Italy, July. Association for Computational Linguistics.

Zweig, G. and Burges, C. J. (2012). A Challenge Set for Advancing Language Modeling. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, pages 29–36, Montréal, Canada, June. Association for Computational Linguistics.