Validating Bundled Gap Filling – Empirical Evidence for Ambiguity Reduction and Language Proficiency Testing Capabilities

Niklas Meyer
Language Technology Lab
University of Duisburg Essen
Duisburg, Germany

Michael Wojatzki
Language Technology Lab
University of Duisburg Essen
Duisburg, Germany

Torsten Zesch
Language Technology Lab
University of Duisburg Essen
Duisburg, Germany

Abstract

Bundled gap filling exercises (Wojatzki et al., 2016) were recently introduced as a promising new exercise type to complement or even replace single gap-fill tasks. However, it is not yet confirmed that the applied creation method works properly and it is still to be investigated if bundled gap-fill tests are a suitable method for assessing language proficiency. In this paper, we address both issues by varying the construction methods and by conducting a user study with 75 participants in which we also measure externally validated language proficiency. We find that the originally proposed way to construct bundles is indeed minimizing their ambiguity, but that further investigation is needed to determine which aspects of language proficiency they are actually measuring.

1 Introduction

Gap filling tasks, also known as cloze tests (Taylor, 1953), are a frequently used for language learning and proficiency testing. The test taker is asked to restore a word that has been omitted from a text or sentence. However, people involved in designing and scoring gap-fill tests are frequently confronted with two major problems: ambiguity and lack of automatability. Ambiguity means that in traditional gap-fill tests frequently more than one word can be used for a gap (Chavez-Oller et al., 1985). For example, the gap in *The kids have to ___ their own lunch* could be filled with *make*, *bring*, *prepare*, or *eat*. However, this fact is often not taken into consideration when it comes to scoring and only one solution is scored as correct. This can lead to high error rates, even with native speakers (Klein-Braley and Raatz, 1982).

Alternatively, there are approaches which allow a set of acceptable solutions, which can improve the validity of gap-fill tests in terms of higher correlations to other tests that measure language proficiency (Brown, 1980). However, this comes at the cost of a higher manual workload and higher subjectivity. An extension of this idea is to weigh the words according to their occurrence in the solutions of participants (Darnell, 1968). However, it could be shown that this scoring procedure has a negative impact on the validity (Brown, 1980).

A way to address these problems is the use of multiple answers, usually the correct solution along with three distractors. The distractors can, however, heavily influence the difficulty of the task. Additionally, using distractors changes the nature of the task from producing a solution to recognizing a solution (Wesche and Paribakht, 1994).

Wojatzki et al. (2016) have recently introduced *bundled gap filling* as an alternative form of gap-filling exercises with a set of gaps in several different sentences, all hiding the same single word. In such an exercise, the learner is confronted with all of the gaps in a bundle at the same time and asked to find the single word to restore all of them correctly. Figure 1 shows examples for all three types of exercises. Wojatzki et al. (2016) showed that the generated bundles decrease ambiguity, but it is still unclear whether the ambiguity reduction was due to their selection procedure or whether any selection of bundled sentences would achieve the same result. Another issue is that in the user study by Wojatzki et al. (2016) all participants had a very high language proficiency level which leaves the question how well bundles work for less proficient learners.

To further investigate these issues, we conducted a user study aimed at comparing the effectiveness

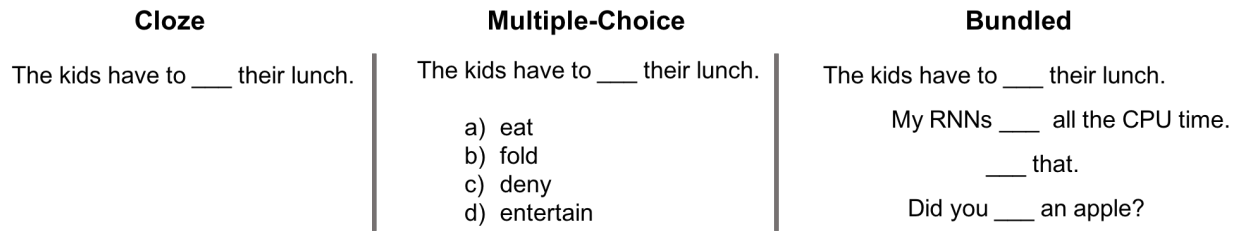

Figure 1: Comparison of exercise types.

of different strategies for computing bundles. In addition, we investigated the relationship between the proficiency level of the test takers and ability to correctly solve bundled gaps. We find that the bundle creation algorithm used by Wojatzki et al. (2016) is disambiguating bundles with a much higher accuracy compared to selecting sentences by chance, while under both conditions the difference to maximally ambiguous bundles is quite high. We also find that the ability to solve bundled gap-fill tasks is indeed substantially correlated ($r = .48$) with the language proficiency of the test takers as measured by *cTest* scores (Klein-Braley and Raatz, 1982). However, the far from perfect correlation implies that further investigation is needed in order to clarify which aspects of language proficiency is measured by bundled gap-fill tests.

2 Bundled Gap-Fill Exercises

In this section, we describe the principle behind bundled gap-fill exercises in order to locate the part of the algorithm that we wish to further validate.

The construction starts with selecting a target word with the surrounding context, i.e. usually a sentence. Depending on the type of exercise or test to be generated the sentence can be taken from a reading assignment, can be provided by a teacher, or can also be a random sentence containing the target word. The algorithm then iteratively adds more sentences to the bundle that contain the same target word. In each iteration the one sentence is selected that maximizes the probability of the target as gap filler for the whole bundle. For the purpose of validating this selection, we propose to select sentences at random and sentences that minimize the probability as competing strategies. We closely replicate the setup by Wojatzki et al. (2016) in our study in order to maximize comparability with their results.

Probability of Gap Fillers We compute the probability of a word fitting the gap using an n-gram language model trained over the two billion word *ukWaC English Web Corpus* (Baroni et al., 2009). We utilize FASTSUBS (Yuret, 2012) with *additive smoothing* (Chen and Goodman, 1999) for efficiently computing the probabilities.

Sentence Base & Target Words We use the GUM corpus (Zeldes, 2016) to select bundle sentences, and we also rely on the same target words as in the original study: four adjectives (*new*, *best*, *full*, *final*), four nouns (*people*, *language*, *information*, *room*), and four verbs (*make*, *want*, *add*, *give*).

Bundle Construction In order to define a target function for unambiguous bundles, Wojatzki et al. (2016) defined the disambiguation level $D(b)$ of a gap bundle b as the log of the ratio between the probability of the target word t and the probability of the most likely word w other than t:

$$D(b) = log \frac{P(F(b) = t)}{\max\limits_{w \in V \setminus \{t\}} P(F(b) = w)}$$

The greater this ratio, the more probable is the target word compared to any other word, and the gap bundle can thus be considered less ambiguous. This mechanism is exemplified in Figure 2.

Given this setup, a bundle for a certain sentence containing the target word is constructed by finding another sentence that contains the target word and which maximizes $D(b)$ for the whole bundle:

$$g_{i+1} = \arg\max_{g \in G_t \setminus b_i} (D(b_i \cup g)), \quad (1)$$

where G is the sentence base and G_t is the set of gaps in G hiding the target word t.

We call this original strategy MAXIMIZE as it maximizes the disambiguation metric $D(b)$. Only

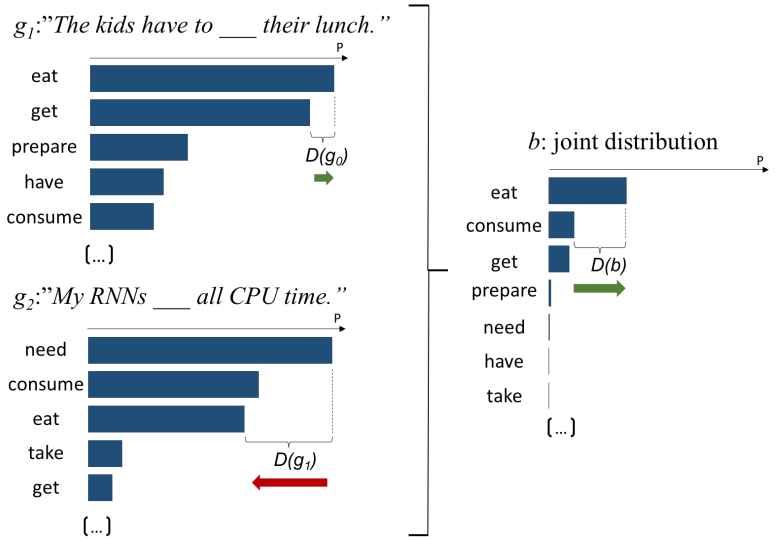

Figure 2: Two cloze tests for the target word *eat* are combined into a bundled cloze test. The diagram illustrates log probability of the possible solutions and how the disambiguation measure $D(b)$ is improved when calculated over the joint distribution.

testing this strategy might hide the fact that randomly selecting sentences with the target word are also likely to increase the disambiguation level. Therefore, we introduce a RANDOM configuration, in which we randomly select sentences. To get better insights into the range of values that the disambiguation level can fall into, we introduce another configuration called MINIMIZE where we change $\arg\max$ to $\arg\min$ in equation 1.

3 Experimental Setup

Given this setup, we can formulate the following research hypotheses:

1. RANDOM Using randomly created bundles results in more ambiguous bundles compared with the original MAXIMIZE setup.

2. MINIMIZE Using bundles that minimize $D(b)$ will lead to even more ambiguous bundles.

Additionally, we are interested in the influence of the language proficiency level of test takers on the success rate in the bundles. We assume that there will be an effect that shows that higher scores are obtained by people with greater proficiency in the English language. We hope to show that the scores in bundled gap-fill tests correlate highly with scores in other language tests, such as the *cTest*. We can thus formulate a third hypothesis:

3. PROFICIENCY There is a high correlation between a test taker's language proficiency and the score obtained when solving gap bundles.

3.1 User Study

To test our hypotheses, we conducted a user study. The study was taken by 118 people of which 75 fully completed the study (52 female, 1 not specified/other gender). As we have three conditions (MAXIMIZE, MINIMIZE, RANDOM, there are 25 participants per condition. The average age of the participants was 22.8 ($SD = 6.9$, ranging from 19 to 67 years). Most of the participants were university students currently enrolled at University of Duisburg-Essen. Additionally, the language proficiency of the participants was measured using a *cTest* that had to be solved after the bundles. For that purpose, we used a *cTest* constructed by the language teaching department of our university.

Participants were shown bundles with an increasing number of sentences. They first saw one sentence with the target word to be restored, then a second, then a third, then a fourth. After each sentence, they were asked to type in the word that (best) suits the gap(s).

Since the GUM corpus is a comparatively small corpus, there are few sentences containing rare words and thus few possible combinations of these sentences. Hence, from the 12 target words used

by Wojatzki et al. (2016), we excluded *room* and *give*, as the bundles in all three experimental conditions were almost identical. Note that in future experiments, this problem could be solved by using a larger corpus from which the bundle sentences are selected.

4 Results & Discussion

In the following, we report and discuss the results of our study.

4.1 Bundle Construction

We first compare the different conditions for creating bundles that are tested in our study: MAXIMIZE, RANDOM, and MINIMIZE. For each condition, we measure the success rate after showing 1, 2, 3, or 4 bundle sentences. A detailed overview of the results per bundle is given in Figure 3, while Figure 4 shows the aggregated results.

As the first sentence is the same under all three conditions, we expect the success rate to be almost the same. The achieved results are close enough to argue that the three subgroups of participants are comparable. For larger bundle sizes, we observe that MAXIMIZE works best, MINIMIZE establishes a lower-bound, and RANDOM is somewhere in between. This shows that the utilized disambiguation measure is able to lower or increase the ambiguity of a bundle (although we usually only want to lower it). How well the RANDOM strategy is going to work largely depends on the properties of the underlying sentence base. If it contains a lot of similar contexts, the success rate might be much closer to the MINIMIZE condition.

Because MAXIMIZE is the same strategy for constructing bundles as was used by Wojatzki et al. (2016), we can compare our results with theirs. However, in their study, all participants had a very high proficiency level while this study was open to participants with different English levels. This explains why our success rates are in general a bit lower, but with the same trend of rising success rates from 1 to 4 sentences in the bundle. In our study the average success rate increases from .10 after only seeing the first sentence to .52 after the fourth. This is a close replication of the numbers from the original study where the increase was from .27 to .78.

Statistical Significance In order to test whether these differences are real differences and not statistical noise, we statistically test our hypotheses. We look at the overall success rates per participant after seeing all four sentences, and conduct a one-way analysis of variance (ANOVA), which indeed confirms both, the MAXIMIZE and the RANDOM hypothesis ($F(2, 72) = 8.93$, $p < .001$). The differences after seeing only one sentence are not statistically significant ($p = .251$). In order to determine which conditions have significantly different arithmetic means, the two a-posteriori tests Scheffé (1953) and Tukey-HSD (Tukey, 1949) were used.[1] Both tests were significant for both combinations (MAXIMIZE, MINIMIZE: Tukey-HSD and Scheffé $p < .001$) and (MAXIMIZE, RANDOM: Tukey-HSD $p = .027$, Scheffé $p = .036$), which further confirms both research hypotheses.

4.2 Language Proficiency

As we have measured the language proficiency of participants using a *cTest*, we can correlate the *cTest* score with the bundle score (of the MAXMIZE condition) to examine whether bundled gap-fill exercises actually measure language proficiency. Figure 5 shows the corresponding scatterplot. The resulting Pearson correlation is $r = .48$. This shows that bundled gap-fill exercises can be used to measure language proficiency, but that both tests seem to measure slightly different constructs. Further research is needed to find out which aspects of language proficiency are actually measured by bundled gap filling exercises, and how bundles relate to other established testing methods.

5 Future Work

Since bundled gap filling is a very recent paradigm, there are various possibilities to deepen the understanding and the validation of the approach. In general, we see three major strands of future research: (i) an refinement of the approach itself, (ii) determining more influencing factors, and (iii) broadening the empirical evidence.

[1] An ANOVA can only implicate that there are generally differences, but is unable to determine which versions show significant differences. Scheffé and Tukey-HSD are the most frequently used post-hoc tests with Scheffé being considered very conservative in contrast to Tukey-HSD.

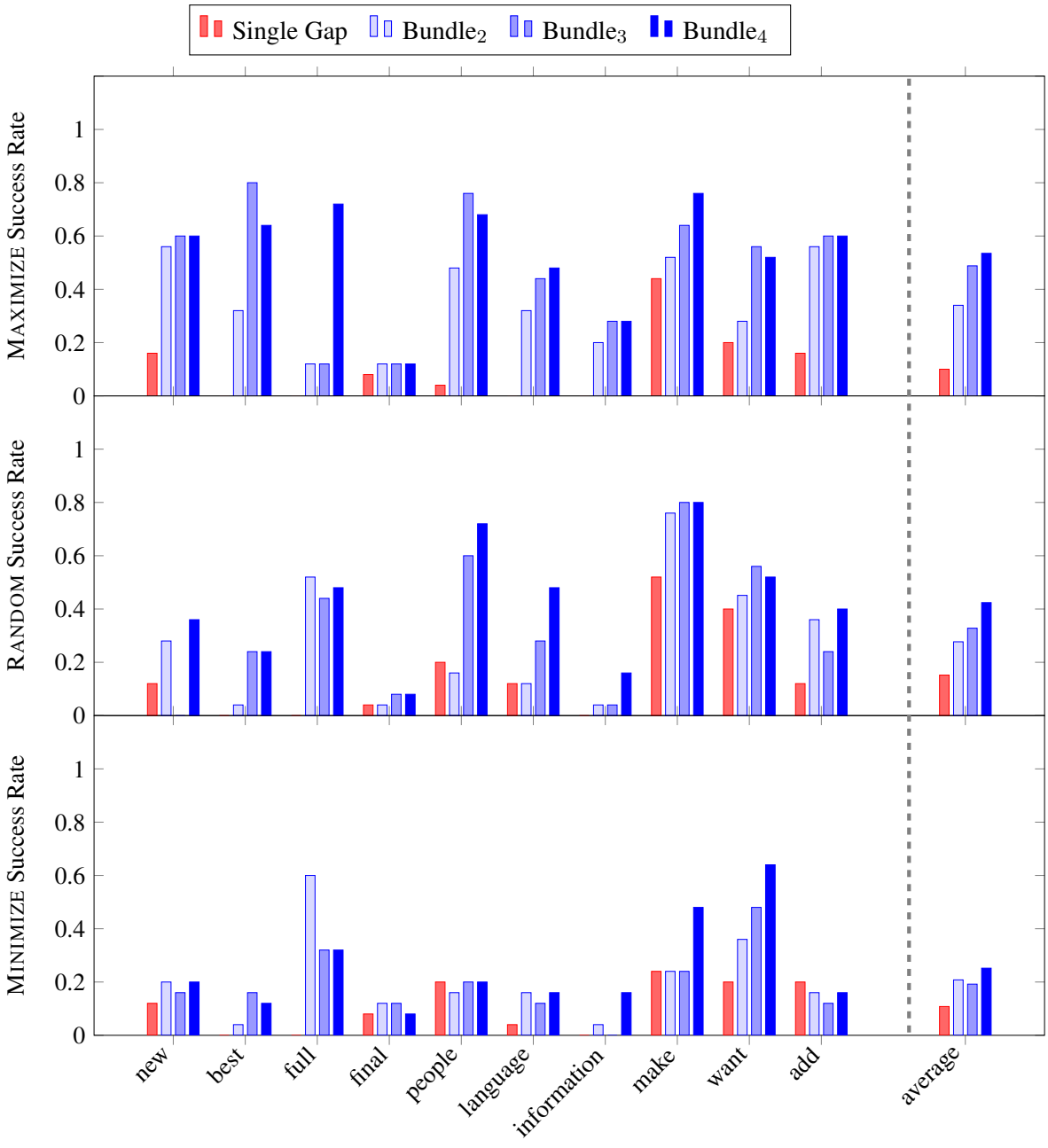

Figure 3: Success rate per item

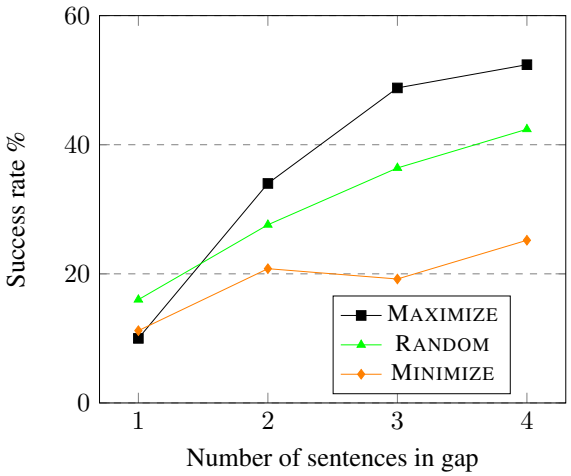

Figure 4: Comparison of strategies for creating bundles

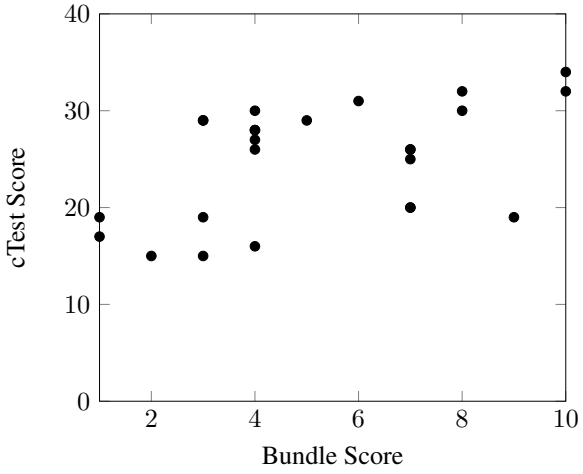

Figure 5: Influence of language proficiency on bundle scores

Refinement The approach for creating bundles could be improved along different lines. First, a different, larger corpus should be used which we expect to lead to even better bundles. Recall, that in the present study, we had to omit two target words which could have been avoided by using a larger corpus. Second, the probabilities of gap fillers have been estimated with a count-based language model. By nature, the used 5-gram model cannot incorporate a context bigger than four words around the gap. However, longer dependencies may indeed play a role when solving gap-fill tests (Bachman, 1982; Chihara et al., 1977). Consequently, future research should clarify whether more advanced language models which are capable of modeling long range dependencies result in even better bundles.

Influencing Factors A number of properties were found to influence the difficulty of gap-fill tests. As bundled gap filling is based on regular gap-fill tests, in future work it should be clarified whether the identified factors also affect the bundled version. The following properties have been shown to have an effect on the difficulty of gaps: Brown (1989) shows that the position of the gap in the sentence and the readability of the passage have an influence on the difficulty of the exercise. Characteristics of the omitted word that affect the difficulty are the length of the word (Abraham and Chapelle, 1992), whether the word is a function word or a content word (Kobayashi, 2002), the frequency of the word in the language (Kobayashi, 2002), and the word origin (Brown, 1989). Consequently, in future work, the set of target words should be systematically varied with respect to the mentioned factors.

Broadening Empirical Evidence In order to strengthen the empirical evidence, future work should aim at creating larger data sets which are closer to existing language learning or testing scenarios. For example, it should be investigated how bundles relate to other state-of-the-art language proficiency tests. For this purpose, bundles need to be introduced to a broader audience and to be integrated into official testing methods. This can help to generate an extensive amount of new data that can further verify bundled gap-filling and show their usefulness in real life scenarios compared to other testing methods. Furthermore, it would be interesting to see how

well results could be reproduced for languages other than English.

The presented results may be biased by the small sample size of this study. Therefore, to further investigate bundled gap-filling and its differences to the *cTest*, it seems necessary to increase the number of test takers.

Last but not least, bundles are also a promising tool for language learning. However, before bringing bundled gap-filling to the classroom, the underlying implementation needs to be taken from prototype to production status. We are currently working on an improved version that we plan to make publicly available.

6 Conclusion

In this work, we have presented an empirical evaluation of *bundled gap filling* (Wojatzki et al., 2016). We confirm that the paradigm is capable of significantly reducing ambiguity in gap-fill exercises – a major problem of this popular exercise type. Moreover, we provide evidence that the originally proposed algorithm for creating bundles is well functioning. As bundled gap-fill scores only moderately correlate with the language proficiency of the participants as measured by a *cTest*, further research is required to determine the properties of bundles.

Acknowledgments

This work was partially supported by the Deutsche Forschungsgemeinschaft (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media". We thank the reviewers for their valuable and encouraging feedback.

References

Roberta G. Abraham and Carol A. Chapelle. 1992. The meaning of cloze test scores: An item difficulty perspective. *The Modern Language Journal*, 76(4):468–479.

Lyle F Bachman. 1982. The trait structure of cloze test scores. *Tesol Quarterly*, pages 61–70.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The {WaCky} wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

James Dean Brown. 1980. Relative merits of four methods for scoring cloze tests. *The Modern Language Journal*, 64(3):311–317.

James D. Brown. 1989. Cloze Item Difficulty. *JALT Journal*, 11(1):46–67.

Mary Anne Chavez-Oller, Tetsuro Chihara, Kelley A. Weaver, and John W. Oller. 1985. When are cloze items sensitive to constraints across sentences? *Language Learning*, 35(2):181–206.

Stanley F. Chen and Joshua Goodman. 1999. An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech & Language*, 13(4):359–393.

Tetsuro Chihara, John Oller, Kelley Weaver, and Mary Anne Chavez-Oller. 1977. Are cloze items sensitive to constraints across sentences? *Language learning*, 27(1):63–70.

Donald K Darnell. 1968. The development of an english language proficiency test of foreign students, using a clozentropy procedure. final report.

Christine Klein-Braley and Ulrich Raatz. 1982. Der C-Test: ein neuer Ansatz zur Messung allgemeiner Sprachbeherrschung. *AKS-Rundbrief*, 4:23–37.

Miyoko Kobayashi. 2002. Cloze tests revisited: Exploring item characteristics with special attention to scoring methods. *The Modern Language Journal*, 86(4):571–586.

Henry Scheffé. 1953. A method for judging all contrasts in the analysis of variance. *Biometrika*, 40(1-2):87–110.

Wilson L. Taylor. 1953. "Cloze Procedure": A New Tool For Measuring Readability. *Journalism Quarterly*, 30(4):415–433.

John W Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, pages 99–114.

Marjorie Wesche and Sima T. Paribakht. 1994. Enhancing Vocabulary Acquisition through Reading: A Hierarchy of Text-Related Exercise Types. Paper presented at the AAAL '94 Conference.

Michael Wojatzki, Oren Melamud, and Torsten Zesch. 2016. Bundled gap filling: A new paradigm for unambiguous cloze exercises. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 172–181, San Diego, CA, June. Association for Computational Linguistics.

Deniz Yuret. 2012. FASTSUBS: An efficient and exact procedure for finding the most likely lexical substitutes based on an n-gram language model. *Signal Processing Letters, IEEE*, 19(11):725–728.

Amir Zeldes. 2016. The GUM Corpus: Creating Multilayer Resources in the Classroom. *Language Resources and Evaluation*, pages 1–32.

Appendix

These are the bundles that were used in the study for the MAXIMIZE condition. All conditions had the same first sentences. Participants were first confronted with only the first sentence, then the second, third and finally the fourth. In total, there were ten bundles.

add

1. Try not to make your words sound like utter and complete gibberish just _____ a little extra than our regular English language.
2. Put the cider vinegar into a small bowl and _____ the soy milk .
3. _____ the wet ingredients to the dry ingredients and beat together (by hand or with an electric hand-held mixer).
4. Here 's a great vegan cupcake recipe to use as a base for whatever flavored icing you want to _____to it.

best

1. They followed _____ practices for anatomical preservation.
2. Just east of Broadway and continuing north and south is Oakland's famous Chinatown, and that to get the real essence of "Chinatown," Oakland rather than San Francisco is your _____ bet.
3. To this day, about 10 or 12 of these World War II Japanese shipwrecks comprise what is considered one of the _____ dive sites in the world.
4. Here 's a great vegan cupcake recipe to use as a base for whatever flavored icing you want to _____to it.

final

1. Not all were pleased with the _____ choice of locations.
2. A _____ thought.
3. The stampede at Islam's most holy site happened at Jamarat Bridge, during an event where pebbles are thrown at a pillar to represent the stoning of Satan as part of the _____ rites of the Hajj.
4. Many people choose to leave out the green, which is lime if you're using original Skittles, and purple, which is grape in the original style, as they can create a weird taste combination or a less than appealing color for the _____ product.

information

1. First, people around the world are desperate for high quality how-to _____.
2. The city maintains several tourist offices, all of which can offer helpful _____ on accommodation, free maps, and bus connections.
3. I don't have enough _____ to answer this question, one way or the other.
4. The Visitors' Center provides _____ on the role of Fort Lee in the War.

language

1. Make sure that it is a _____ that while speaking, you don't get a literal knot in your tongue!
2. As they design their web pages for the newer browsers with advanced web technology and geared to the newest web core markup _____ HTML 5, they are forced to accommodate older out-of-date technology to support IE6 users.
3. Be fluent in your own made up _____ and start spreading this to your friends, family and strangers!
4. Write your own poem/novel/story with your own made up _____.

make

1. However, paying people to write and edit articles ultimately means that you have to _____ one of two sacrifices.
2. In the 1960s and 1970s, many 19th century neoclassical buildings, often small and private, were demolished to _____ way for office buildings, often designed by great Greek architects.
3. The single most costly thing we spend on is rent and advertising, those two together _____ up the bulk of what we spend.
4. You should _____ sure that your clothing covers at least your shoulders and your knees and some places may require that you wear ankle-length pants or skirts and long sleeved tops.

new

1. We took quite a few _____ girls over there back then in 2005, leading into the World Cup in the Netherlands.
2. Athens today is ever evolving, forging a brand _____ identity for the 21st century.
3. The Museum of Flight in Seattle, Washington was also proposed as another location for a shuttle, going so far as to build a _____ building to house an orbiter.
4. In March, a bundle of blueprints for a _____ headquarters for the military's counterterrorism unit were found stuffed in the trash on a downtown street.

people

1. It emphasizes consumerism, the belief that success always goes to _____ who merit it due to their abilities, dedication and qualifications, and reinforces, rather than changes, existing ideas related to gender, ethnicity and nationality.
2. On the other hand, this isn't to say that you should necessarily make jokes at other _____'s expense, as this can make you seem mean and petty.
3. Telling good jokes is an art that comes naturally to some _____, but for others it takes practice and hard work.
4. Moreover, electing a third-party governor represents a repudiation of politics as usual, and the major party legislators will face changed constraints and incentives, meaning that much more is possible than many _____ assume, especially with strong leadership.

want

1. Why did she so badly _____ to attend?
2. For instance, you might say something like: "If you like those guys, you might _____ to check out this band called Manic Albatross - they're like the Beatles, only darker.
3. How do you approach the difficult challenge of talking to the Palestinians when, in the end, they dont _____ Israel to exist.
4. "We _____ to thank all of the locations that expressed an interest in one of these national treasures," said Bolden to the gathered crowd which contained many KSC employees.

full

1. However, the _____ fuselage trainer, that every astronaut including [former Museum of Flight CEO] Bonnie Dunbar has been trained on, will soon call the Museum of Flight home.
2. Another thing non-locals don't often realize is that Cleveland's long history of industrial wealth has left it chock _____ of cultural riches as well as the beginnings of a "sustainable city" movement.
3. If you buy too many boxes you can return the unused for a _____ refund.
4. York is _____ of magic and a wonderful place to bring children!