In D. Reitter & F. E. Ritter (Eds.), Proceedings of the 14th International Conference on Cognitive Modeling (ICCM 2016). University Park, PA: Penn State.

Investigating and Simulating the Effect of Word Fragments as Orthographic Clues in Crossword Solutions

Kejkaew Thanasuan (kejkaew.tha@kmutt.ac.th)
Learning Institute, King Mongkut's University of Technology Thonburi
Bangkok, 10140 Thailand

Shane T. Mueller (shanem@mtu.edu)
Department of Cognitive and Learning Sciences, Michigan Technological University
Houghton, MI 49931 USA

Abstract

A number of models of word structure represent orthography in terms of features indexing individual letters and adjacent letter pairs within the word. This permits word parts to be represented independent of position, but leaves the open question of whether partial letters arranged in multiple clusters (fragments) provide better memory retrieval cues. To answer this, we conducted a study in which expert and novice crossword players completed crossword problems for words of different lengths and with different numbers of letter cues. Although expertise, word length and a number of cues provided strong predictors of accuracy and response times, within each cue/word-length condition, neither the number of word fragments nor the maximum size of word fragment provided a consistent advantage. A computational model using letter pairs as features, but no higher-order representation of orthography, accounted for effects of expertise, word length, and number of cues, and similarly did not produce systematic effects on the number or sizes of fragments. Results suggest that, to a first approximation, letter-pair representations are sufficient to account for the performance in word stem and word fragment completion, crossword, and potentially other word reading/identification tasks.

Keywords: crossword expertise; memory retrieval; orthographic clue; crossword recognitional-based decision-making model

Introduction

Models of reading and word representation in Latin-character languages have often represented orthography in terms of features associated with both individual letters and *letter pairs* (Grainger & Van Heuven, 2003; Thanasuan & Mueller, 2014). For example, the feature-based representation for FORCE would include F, O, R, C, and E, along with *F, FO, OR, RC, CE, and E* (where * indicates a word boundary). This scheme permits representing a sequence without coding the absolute position of a letter within a word, and so it enables similarity-based comparisons to be robust to prefixes and suffixes (ENFORCE, FORCEFIELD) with graded similarity to some grammatical modifications (FORCING). Such models have been successful at accounting for data in a number of reading, memory, and word completion paradigms, but less is known about whether higher-order orthographic representations such as trigrams (or "Wickelphones") are useful, and whether cues involving multiple letter pairs are better than those involving fewer pairs.

Although such representations have often been tested in a general population of readers (who are assumed to have substantial experience with the problems of memory retrieval based on orthographic information), another factor to consider is whether extensive deliberate practice with word-fragment completion changes the level of representation used, or permits better use of multiple word fragments in cueing a correct word. Thus, expert word game players may produce fundamentally different results, showing evidence of different or higher-order representations in word completion tasks in their domains of expertise.

A number of researchers have examined crossword solvers to identify their memory retrieval and problem-solving abilities and strategies. For example, Nickerson (1977, 2011) explored crossword puzzle solving processes relating to lexical memory and categorized daily crossword cues in order to understand information retrieval of the solvers. Moreover, Toma, Halpern, and Berger (2014) compared visuospatial and verbal working memory among college students, crossword and Scrabble experts using a symmetry span task and a reading span task. They found that participants from the two elite groups performed the cognitive tasks better than the novice group, but the results between the two groups were not statistically different.

Mueller and Thanasuan (2013); Thanasuan and Mueller (2014) developed and implemented computational models of crossword solving based on the Recognitional-primed decision making model (RPD; Klein, 1993), the Bayesian Recognitional Decision Model (BRDM; Mueller, 2009), and a computational model of word-stem completion (Mueller & Thanasuan, 2014). The models used data from a database of millions of real crossword clues and answers, and used a letter-pair feature set to represent orthography. Expertise effects were accounted for in terms of speed, strategy, and retrieval fluency, and although the models performed better than novices, they did not achieve as good performance as did experts. This may have arisen in part because of the orthographic representation; a computer performing logical template-based matches of word stems can generally reduce the candidate set substantially more than our representation (see Ginsberg, 2011), which would have improved performance significantly.

Letter Clusters as Orthographic Clues

Typically, a partially-filled answer in a crossword grid is easier to solve than one with no letter cues, in part because it limits the search set in mental lexicon (Nickerson, 2011), and

provides additional cues for retrieval. Thanasuan and Mueller (2014) concluded that although crossword experts used semantic information as a primary constraint, they also relied on orthographic knowledge and visual pattern recognition to complete the puzzles. Supporting this, Mueller and Thanasuan (2013) found that when crossword experts were presented with easy word-stems (i.e. three or fewer missing letters), the accuracy was about 80%, whereas novices were able to complete them correctly only about 40%.

However, those studies did not examine whether clusters of two or more letters were more helpful in memory retrieval than if the same letters are dispersed throughout the clue. Other tasks, such as the cue-facilitated retrieval paradigm, have been used to investigate the role of letter clusters in word completion. For example, Horowitz, White, and Atwood (1968) used this paradigm to determine whether the type of letter cluster (the first, middle or last three-letter clusters of a word) impacted the ability to recall nine-letter words, and found that the first three-letter fragment was the most helpful. Likewise, Dolinsky (1973) compared the cue retrieval process using syllabic and non-syllabic letter clusters as cues. They found that the middle syllabic units were very helpful on word retrieval, but the syllabic clues did not facilitate the recall performance more than the non-syllable fragments. In addition, Goldblum and Frost (1988) studied a cue facilitation effect of letter clusters using a crossword paradigm task. They hypothesized that different structures of sub-lexical units, which include syllable, pronounceable non-syllable, unpronounceable cluster, and nonadjacent letters, might differently influence word retrieval. They found that the small units of syllabic fragments were the best retrieval cue. They also found that the syllables with phonological units such as "- -SEP- - - - -" of a target word "INSEPARABLE" assisted the solvers more than morphemic units (i.e., units linked to meaning) such as "- - - -PAR- - - -" of the same target answer. This suggests that the position of letter clusters within a word and within syllable boundaries may be important.

We suggest that two properties regarding the usefulness of letter clusters are not well understood. The first is whether, for a fixed number of letters in the cue, does their arrangement into clusters impact accuracy of retrieval? For example, if four letters are given for the word "HOUSECAT", they might be "H-U-E-A-", "HO- -EC- -", or "HOUS- - - -". The first has no clusters (sets of adjacent letter pairs), the second has two clusters, and the third has one cluster. It may be that, for either novices or experts, a better or a worse cue is provided when letters are arranged into more clusters. In terms of models of representation, if an advantage exists, this may indicate the use of higher-order representations and require adapting existing models to account for such results. The second property is the maximum length of the cluster—do larger clusters form better cues than smaller clusters? Regarding the previous example, the largest cluster size is one, two and four, respectively. These may differently impact solution probability.

Crossword Solving Model

Previously, we have described a cognitive computational model of crossword solving that accounts for expertise, word length, clue difficulty, and letter-clue effects (Mueller & Thanasuan, 2013; Thanasuan & Mueller, 2014), that we will use in this study. The model simulates crossword answers via two independent routes: semantic and orthographic memory associations (see Figure 1). Mueller and Thanasuan (2013) indicated that the best model representing crossword solving performance was the dual route model with three different conditions for novice, expert and best performance simulations. This model first attempts retrieval via the orthographic route. If it fails, the model searches via semantic associations.

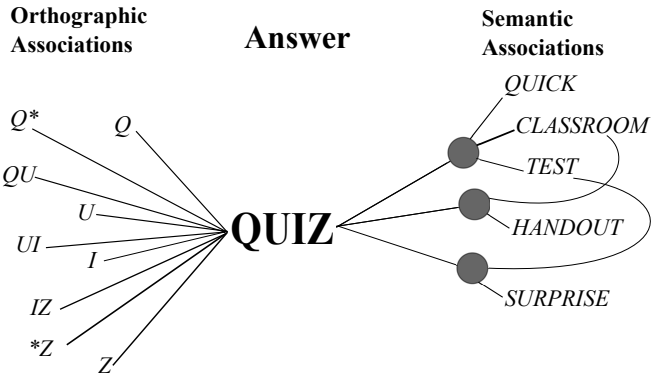

Figure 1: Example of semantic and orthographic routes

The orthographic route model works by representing orthography according to associations between features of a word and crossword answers. Orthographic features that were used include letters, letter pairs, and a distributed length code that helps limit possible answers to words of similar length to the clue. The model was trained on a lexicon of more than 4 million crossword clue answers, and so it has rich associations between these features and existing crossword answers. The representation does not include any higher-order letter clusters (sequences of three or more letters). If there are word fragment effects (depending on size or number of clusters) that cannot be accounted for by the model, this may indicate higher-order representations are needed. The challenge for higher-order representations is that the number of features required scales with the power of the cluster size, making naive representations unmanageable, especially when most of these features never appear in a given language. The alternative would be to begin incorporating syllable representations informed by phonology (see Fudge, 1969; Mueller, Seymour, Kieras, & Meyer, 2003), morphology, or information-theoretic measures, the present study seeks to determine whether this increased complexity is necessary.

To test the model, we investigated the role that word fragments play in crossword puzzle solutions among both experts and novices. A crossword paradigm task was used in which participants were given single clues with partial

letter hints. Although the task requires matching or retrieval from both semantic memory and orthographic memory, we focused on studying the solving mechanisms associated with orthographic clues and word fragments. Furthermore, to assess the role of word fragments on representation, a crossword-solving model based on Mueller and Thanasuan (2013); Thanasuan and Mueller (2014) was adapted to simulate crossword answers and response times. Our approach is to examine whether factors that would be consistent with the use of higher-order representations improve performance, and to demonstrate whether the model shows a similar effect. To do this, we examined whether the size of the largest cluster, and the number of clusters in a clue had an impact in retrieval times or accuracy for both novices and experts.

Experiment

Participants

Eighty-five undergraduate students were recruited from Michigan Technological University (MTU) subject pool as crossword novices. In addition, 113 crossword experts were recruited from online crossword communities. Both groups of participants completed an online crossword study and a demographic survey via web browser. The study protocol was reviewed and approved by MTU Institutional Review Board.

The survey was given to participants at the beginning of the experiment. The novices were 19.94 $\pm$ 1.5 years old in average. Eighty-three percent of them rarely or never solved crossword puzzles, but some of them played other word games such as Scrabble or Words With Friends. The experts were 45.07 $\pm$ 15.99 years old in average. They reported that they have solved crossword regularly for 16.76 $\pm$ 15.35 years and 44 percent of them have participated in crossword tournaments such as American Crossword Puzzle Tournament (ACPT).

Materials and stimuli

Six sets of 15 crossword paradigm task problems were given to participants in this study. The answers were limited to only a set of four, six and eight letter words. In each trial, we gave participants a crossword clue along with some randomly filled letters, as shown in Figure 2. The number of present letters ranged from zero (no letter cue) to only one missing letter (almost a complete answer). They had 15 seconds to solve each trial and only one chance to give an answer.

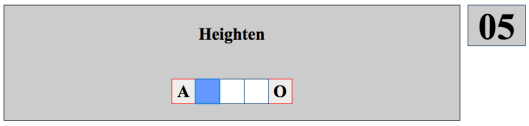

Figure 2: Example of the crossword paradigm task

Experimental Results and Model Simulation

Data from 198 participants were analyzed in this study. Accuracy was assessed by whether the final set of letters were exactly correct after the enter key was pressed to confirm the response, whereas response times were measured from a starting time (when a participant first saw a trial) until the participants hit the enter key. Average accuracy and response time of the crossword experts were 0.76 $\pm$ 0.19 (67 from 90 trials) and 4.82 $\pm$ 2.67 seconds per each trial, respectively. Meanwhile, a mean of success rate of the novices was 0.53 $\pm$ 0.15 (47 from 90 trials) and an average response time was 6.69 $\pm$ 1.88 seconds per each trial. Retrieval times of both experts and novices were estimated from the starting time until the first key was pressed (when the participants typed a first letter to an answer space). An average retrieval time of the novices was 5.73 $\pm$ 1.56 seconds, so the time for the simulations of the Novice model was 0.57 (5.73 divided by a search set size of 10). The time of the experts was 3.13 $\pm$ 1.01 seconds, then the time for the simulation of the Expert model was 0.22 (3.13 divided by a search set size of 25). Average typing speeds of each keystroke of the novices and the experts were 0.35 seconds and 0.22 seconds, respectively.

Figure 3 shows solving performance across the number of letter cues for both success rate and response times. It indicates that the experts performed faster and better than the novices did. Also, both groups of participants performed better when the number of present letters increased. A one-way Analysis of Variance (ANOVA) was used to analyze letter cueing effects, which indicated a significant improvement for both success rates and the response times of both experts and novices as the number of letter cues increased (p-value < 0.05).

Model simulation

To simulate data, we compared two model parameter settings per expertise conditions (Nov=Novice, Exp=Expert), varying the search set size (10, 10, 25, and 50 for models of Nov1, Nov2, Exp1 and Exp2, respectively), recovery (0.5, 3, 25 and 100), and retrieval speed (130 ms, 130 ms, 570 ms, 570 ms). We assigned the same value (10^{-9}) to smoothing orthographic and semantic parameters in order to optimize and balance solving performances of these two routes. These parameters increase chances of getting answers that have been associated to only one item in the memory activation distribution. All other parameters were fixed to the same levels used in previous simulations. The search set size parameter impacts the number of highly-active candidates retrieved during solution; the recovery parameter affects the probability that a response can be generated once a memory trace has been selected, and the retrieval speed affects the time needed to retrieve a candidate and verify whether or not it is correct. Retrieval and typing speeds were determined from the experimental results. Reading speed was taken from Ziefle (1998)'s study regarding an effect of display resolution, which is about 0.33 seconds per word. Then, a solving time of each trial was

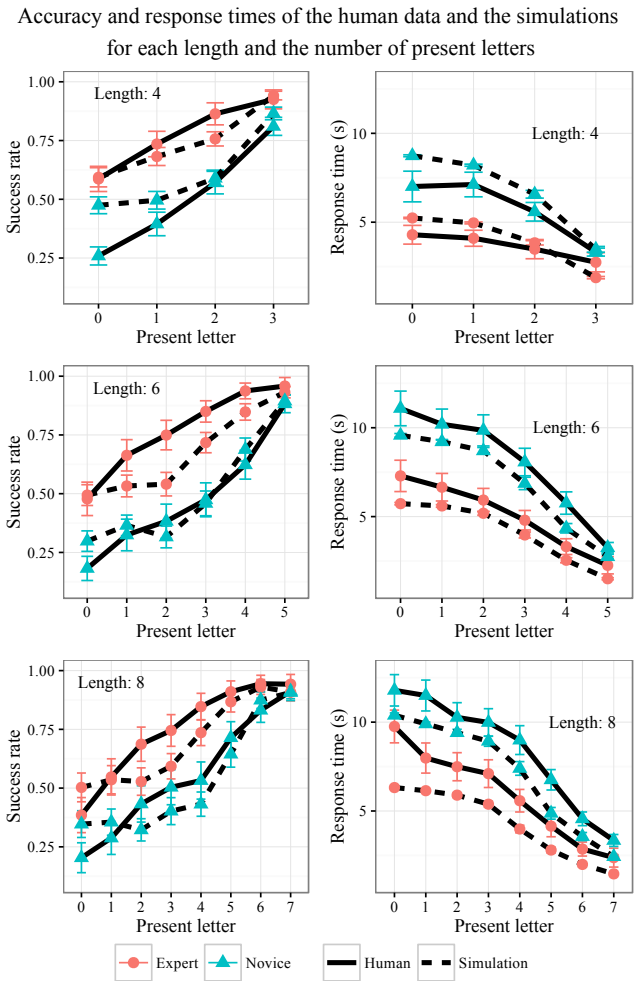

Figure 3: Dots represent means of correct responses and reaction times and error bars indicate 95% confidence intervals

estimated from Equation 1:

$$T_{solving} = cl * t_{reading} + n * t_{retrieval} + wl * t_{typing} \quad (1)$$

where cl is the total number of words in a clue, $t_{reading}$ represents the reading speed, n is the number of candidate answers that the model generates before it gets the first answer that fits the pattern, $t_{retrieval}$ is the retrieval times of the novices and the experts, wl is word length, and t_{typing} is the typing speeds of the novices and the experts, which are 0.35 seconds and 0.22 seconds, respectively. We used two model settings to enable two bracketed models that can account for different levels of expertise.

Table 1 shows simulation results from the crossword play model across the four different settings; two novice models (Nov1 and Nov2) and two expert models (Exp1 and Exp2). The Nov2's success rate and response times were almost the same as the novice data, whereas Exp1 and Exp2 produced success rates and response times closely related to the expert data. Model fits were assessed via Root-Mean-Square Error

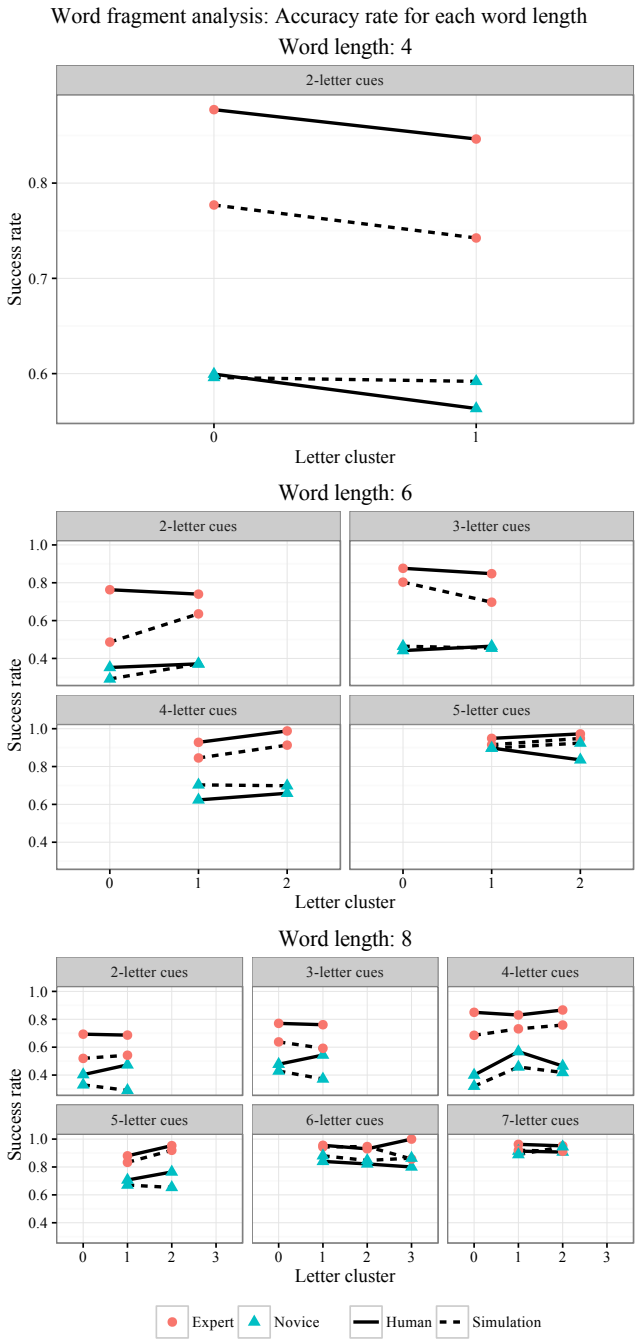

Figure 4: Results for 4, 6, and 8-letter words. Each panel represents a fixed number of letter cues, and the horizontal axis represents the effect as the number of clusters increases.

(RMSE), shown Table 1. The Nov2 model was the best fitting model on both accuracy and response times of the novice data. Meanwhile, Exp2 was the best fitting model in accuracy and Exp1 was the best fitting model in response times of the expert data. However, we chose the Exp1 model to represent the expert data, since an average RMSE of the accuracy and

the response times of Exp1 was less than the other. Moreover, Figure 3 compares the results of Nov2 and Exp1 to the human data.

Table 1: Model results (mean and standard deviation) and Root-Mean-Square Errors (RMSE)

Parameter	Model	Mean (SD)	RMSE	
			Novice	Expert
Acc.	Nov1	0.38 (0.06)	0.17	0.42
	Nov2	0.55 (0.05)	**0.09**	0.26
	Exp1	0.7 (0.04)	0.21	0.1
	Exp2	0.77 (0.04)	0.28	**0.07**
RT (s)	Nov1	4.71(0.46)	1.58	1.9
	Nov2	6.0(0.39)	**1.24**	2.2
	Exp1	3.75 (0.22)	3.83	**1.4**
	Exp2	5.49 (0.37)	2.09	1.83

Note: The bold numbers indicate the smallest value in each performance.

Effects of number of clusters on completion accuracy

The number of letter clusters (groups of two or more adjacent letters) was computed for each orthographic clue. Figure 4 shows means of success rates of each letter cluster for each word length. Each connected line shows how performance changed within each cue number and word length condition as the number of letter clusters increased. Although there were occasional fluctuations, there were no systematic effects of the number of letter clusters on either experts and novices, which was confirmed by a logistic regression and a Chi square goodness-of-fit test (experts: $\chi^2(3) = 6.01, p = .11$ and novices: $\chi^2(3) = 3.22, p = .36$).

The models reasonably replicated human solving abilities (see Figures 3 and 4), although they somewhat underperformed expert performance as a function of number of letters in the cue. On particular word length/number of cue combinations, the model or the humans saw changes with respect to number of clusters and these were sometimes shown in both the model and human data. To the extent that any of these are systematic, they may have stemmed from the way that for any particular word, some combinations of letter clues will do a better job of reducing or eliminating alternative completions, and these combinations may be covered better or worse by clusters of letters. Thus, although the number of letters given improves solution performance significantly, neither the model nor the human data showed systematic effects of this variable on solution accuracy.

Effects of maximum cluster size on completion accuracy

The effects of cluster sizes on accuracy are shown in Figure 5. Similar to the finding with number of clusters, there was little systematic effect of the size of the largest cluster on solution accuracy. Participants occasionally performed better when they were given more adjacent letters, such as for two-letter and four-letter cues of eight-letter words, and in these cases similar effects were seen for both the model and the data. The logistic regression and the Chi-square test were conducted to determine the effects of maximum cluster sizes. The results indicated that the effects were non-significant on accuracy for

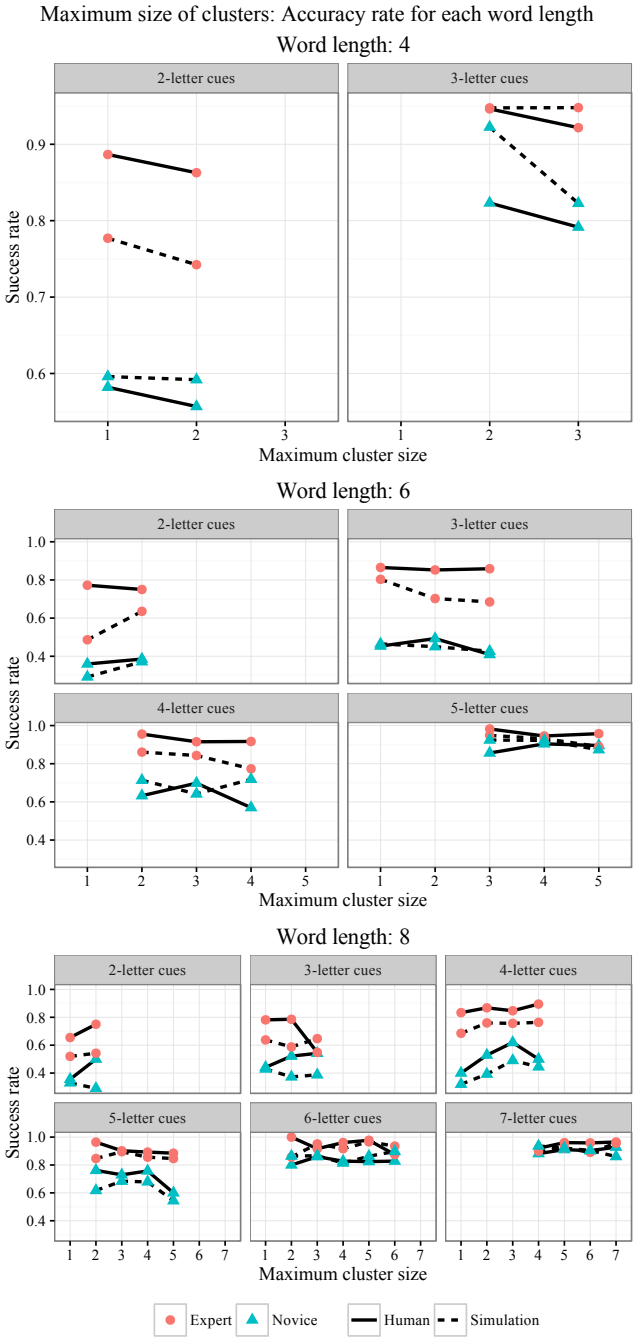

Figure 5: Accuracy by maximum size of cluster. Within each panel (that show a different number of cues) each line shows novice and expert accuracy for different word lengths, as the maximum size of a cluster increases.

novices ($\chi^2(6) = 3.19, p = .78$), and marginally significant for experts ($\chi^2(6) = 12.25, p = .057$). To the extent that the effect exists among experts, it showed that smaller maximum cluster size (which is coupled with greater distribution of isolated letters) tended to lead to better performance, although this was not always the case in each condition.

Discussion

The goals of this study were to investigate the effect of word fragments to determine whether humans enjoyed an advantage of using word fragments that were not predicted by a model using letter-pair representations. We hypothesized that if higher-order orthographic representations were in use, then for a given word length and number of presented letters, when the number or size of word fragments increased, both experts and novices would improve. The findings from the crossword paradigm task suggests that although experts performed crossword solving better than novices did, and the number of letter cues influenced the solving performances on both accuracy and response times, the properties of word fragments we looked at had little impact on performance. Similarly, the model accounted for effects of expertise, word length, and stem size, but showed no systematic effects on these properties of letter clusters. This suggests that, to a first approximation, orthographic models using letter-pairs are still appropriate in representing word retrieval processes.

Nevertheless, we believe that higher-order representations may prove useful in understanding more complex and advanced crossword solving behavior. It may be features associated with morphological, phonological, and syllabically-consistent clusters will both provide substantial advantages for solving, as well as be critical for explaining how long crossword clues are solved. For example, a typical American-style puzzle published in a Saturday New York Times puzzle will have two or more answers that are 15 letters long (ONCEINALIFETIME has been used at least 20 times), and such clues are even more common in cryptic-style crossword puzzles popular outside the United States. More than 5000 such answers have appeared in print[1], and almost all of them are short multi-word phrases. It is likely that the division between composing multiple words, and composing a single word from multiple meaningful lexical units that fit together according to grammatical rules is not as clear as it might seem. Regardless, the representations, processes and mechanisms a model would require to solve multi-word clues (ie., word-level features) would be similar to what would be needed to use syllable or morpheme-level features for single-word clues, and addressing this problem may help understand segmentation in reading, listening, and non-latin languages different rules and practices of segmentation. Consequently, we believe that it may remain useful to consider whether pronounceable clusters, syllables, or morphological units can form features, and to test this in future experiments.

[1] see `http://www.xwordinfo.com/Fifteen`

Acknowledgments

The experiment was conducted while KT was a graduate student at Michigan Technological University.

References

Dolinsky, R. (1973). Word fragments as recall cues: Role of syllables. *Journal of Experimental Psychology*, *97*(2), 272–274.

Fudge, E. C. (1969). Syllables. *Journal of linguistics*, *5*(2), 253–286.

Ginsberg, M. L. (2011). Dr. fill: Crosswords and an implemented solver for singly weighted csps. *Journal of Artificial Intelligence Research*, 851–886.

Goldblum, N., & Frost, R. (1988). The crossword puzzle paradigm: The effectiveness of different word fragments as cues for the retrieval of words. *Memory & Cognition*, *16*(2), 158–166.

Grainger, J., & Van Heuven, W. (2003). Modeling letter position coding in printed word perception. *The mental lexicon*, 1–24.

Horowitz, L. M., White, M. A., & Atwood, D. W. (1968). Word fragments as aids to recall: the organization of a word. *Journal of Exp. Psychology*, *76*(2), 219–226.

Klein, G. A. (1993). A recognition-primed decisions (rpd) model of rapid decision making. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action* (pp. 138–147).

Mueller, S. T. (2009). A bayesian recognitional decision model. *Journal of Cognitive Engineering and Decision Making*, *3*(2), 111–130. doi: 10.1518/155534309x441871

Mueller, S. T., Seymour, T. L., Kieras, D. E., & Meyer, D. E. (2003). Theoretical implications of articulatory duration, phonological similarity, and phonological complexity in verbal working memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *29*(6), 1353–1380.

Mueller, S. T., & Thanasuan, K. (2013). A model of constrained knowledge access in crossword puzzle players. In R. West & T. Stewart (Eds.), *The 2013 international conference on cognitive modeling (iccm12)* (p. 275).

Mueller, S. T., & Thanasuan, K. (2014). Associations and manipulations in the mental lexicon: A model of word-stem completion. *Journal of Mathematical Psychology*, *59*, 30–40.

Nickerson, R. S. (1977). Crossword puzzles and lexical memory. In S. Dornic (Ed.), *Attention and performance vi* (pp. 699–718). Hillsdale, N.J: Lawrence Erlbaum.

Nickerson, R. S. (2011). Five down, absquatulated: Crossword puzzle clues to how the mind works. *Psychonomic Bulletin & Review*, *18*(2), 217–241.

Thanasuan, K., & Mueller, S. T. (2014). Crossword expertise as recognitional decision making: an artificial intelligence approach. *Frontiers in Psychology*, *5*. doi: 10.3389/fpsyg.2014.01018

Toma, M., Halpern, D. F., & Berger, D. E. (2014). Cognitive abilities of elite nationally ranked scrabble and crossword experts. *Applied Cognitive Psychology*, *28*(5), 727–737.

Ziefle, M. (1998). Effects of display resolution on visual performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*(4), 554–568.