Measuring Hint Level in Open Cloze Questions

Juan Pino and **Maxine Eskenazi**
{jmpino, max}@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh PA 15213
USA

Abstract

Providing the first few letters of a missing word in a sentence gives information about this word. This paper attempts to measure the information transmitted in that case. In order to do so, we analyzed response accuracy for *open cloze questions*, that is fill-in-the-blank questions without multiple choice answers. In this study, native and non-native speakers of English answered a series of open cloze questions that were semi-automatically generated. Hints were provided that consisted of the first few letters of the missing word. Results showed that question difficulty, hence the quantity of information transmitted, is related to the number of letters that are provided, to physical properties of these letters and to syllables formed by these letters. Performances did not appear to depend on letter or syllable frequency. Controlling hint level in a word completion task is critical in order to provide practice exercises adapted to student levels.

Introduction

In this paper, we attempt to measure the level of hint provided by the first few letters of a missing word in a fill-in-the-blank question. Our goal is to adapt the question difficulty to student levels in an intelligent tutoring system for vocabulary learning. Providing the first few letters of a missing word in a sentence gives information about this word. This information is encoded in diverse ways. First, letters themselves provide a piece of information. When combined together, letters carry other information unit types. For instance, they provide phones and phone combinations. The frequency of phones and phone combinations in the language considered might also determine the amount of information provided. Finally, letters or phones can be assembled to form syllables, another information unit type. Syllable frequency might also play a role in the amount of information transmitted. This paper attempts to measure the information transmitted when providing the first few letters of a missing word in a sentence by exploring the role of the different information pieces mentioned above. In order to do so, we analyzed response accuracy for *open cloze questions*. Response accuracy directly reflects question difficulty which in turn reveals how much information was provided. We designate this amount of information as *hint level*. Measuring hint level is critical in order to adapt question difficulty to a specific student's level.

Cloze and *open cloze* questions are fill-in-the-blank questions; while the former have multiple choice answers, the latter do not. Cloze questions are currently used in the REAP system (Heilman et al. 2006) as part of its assessment and practice module. REAP is an intelligent tutoring system that teaches vocabulary to English as a Second Language students. It retrieves authentic documents from the World Wide Web and filters them according to several pedagogical constraints such as length, topic and reading difficulty (Collins-Thompson and Callan 2005). In a student session, readings are followed by practice exercises, namely cloze questions, in which the system updates the learner's model to provide suitable documents for the subsequent session. Benefits of cloze questions include providing a reliable measure of vocabulary knowledge and being easy to assess. Currently, the system uses manually generated cloze questions, which is a time consuming process. Pino and colleagues (2008) developed a strategy to generate good quality cloze questions. While the sentences generated were satisfactory, the strategy did not produce distractors, i.e. wrong choices, of sufficient quality.

From a pedagogical point of view, it seems preferable to use multiple choice cloze questions rather than open cloze questions. Indeed, open cloze questions have several different answers very frequently, which can confuse students. Furthermore, answers to open cloze questions are difficult to assess, both by a computer and a human. However, open cloze questions demand more productive knowledge (Nation 2001) from students than cloze questions. Providing hints to open cloze questions in the form of the first letters of the missing words is a way to narrow down the number of possible answers and to make assessment easier, while at the same time still requiring active knowledge from the students. Controlling the amount of hint provided is critical to adapt question difficulty to student levels (Wood and Wood 1999). The REAP system already adapts reading difficulty to student levels. Adapting the difficulty of the questions seems a logical next step in the development of the tutor.

In the following sections, we first justify the use of this particular hint form and ground it in related work. We then describe our strategy to produce open cloze questions. Fi-

nally, we present an experiment attempting to determine hint level. In this experiment, subjects answered a series of open cloze questions. We provided the first few letters of the missing words and varied the number of letters provided in order to make the difficulty level vary. Our analysis relates information provided, or equivalently question difficulty to the number of letters provided, letter properties (vowel, consonant) and syllables.

Related Work

Providing hints for open cloze questions in the form of the first letters of the missing words is not a new idea. Laufer and Nation (1999) used this technique to ensure that only one answer is possible while Klein-Braley and Raatz (1984) systematically provided the first half of the missing word. The aim of this paper is to study the relation between hint provided and response accuracy in order to be able to measure the amount of information provided by hints and adapt hint levels to student levels.

We chose to provide the first letters of the word rather than the last ones or randomly selected letters. Indeed, it has been shown that first letters are more important than last ones for word recognition (Oléron and Danset 1963). Yet it has been argued that letters might not be the sole unit to consider for word apprehension.

Chunks of letters, more precisely letter n-grams, can be considered as a unit. Lima and Inhoff (Lima and Inhoff 1985) have studied the relation between eye movements and trigram frequency and shown that words in which the first letter trigram has a lower *type* frequency have a longer eye fixation than low constraint words.

The role of the syllable has also been studied in word recognition. Carreiras and Perea (2004) have shown that pseudowords with high frequency first syllables have a faster response time than pseudowords with low frequency first syllables, independently of the presence of stress on the first syllable. They also showed that the second syllable has no significant effect on response time.

Their experiment was conducted with native speakers of Spanish. Cutler (1997) also showed that French listeners detect syllables faster than mere letter chunks, although earlier, Cutler et al. (1986) conducted a syllable monitoring task experiment and argued that English listeners rely on phonemes rather than on syllables. Peretz et al. (1998) confirmed this trend for word completion tasks.

This work is based on a different kind of word completion task. In (Peretz, Lussier, and Béland 1998), subjects have to complete words in isolation while the subjects of our experiment complete words *in context*. In this framework, we will study several factors for question difficulty. One obvious factor is the number of letters provided. We also want to explore the role of the letter type – vowel or consonant – in the word completion task. Another factor is the role of syllables. More specifically, we will examine the performance gain obtained when a syllable is provided; we will also investigate the influence of stress or absence of stress on the first syllable. Finally, we want to explore the relation between letter frequency, n-gram letter frequency, syllable frequency and question difficulty.

Open Cloze Question Generation

Hint levels are evaluated in the context of open cloze questions. In this section, we describe a strategy to generate these questions. We want to generate a sentence containing a word w. We first gather several sentences s_1, ..., s_n containing the word w. The sentences are extracted automatically from several online dictionaries. Not all of them are suitable for open cloze questions; we want to keep those which are grammatically correct and define a sufficiently rich context to allow a limited number of possible answers. We measure the richness of the context for sentence s_i using four linguistic criteria: collocations, grammatical complexity, grammaticality and length. Collocation statistics are gathered from a database of documents. We then sum the collocation scores between the target word in s_i and words in a window of five words around the target word. This results in a collocation score for s_i. The grammatical complexity is measured by counting the number of clauses contained in s_i after parsing s_i with the Stanford parser (Klein and Manning 2003). The grammaticality score for s_i corresponds to the log-likelihood of its most probable parse tree output by the Stanford parser. The grammaticality score is normalized with the length of s_i in order to avoid length bias. Finally, the four linguistic scores are manually weighted and linearly combined to output a global score for s_i. We score $s1$, ..., s_n as we just described and rank them. The top sentence is retained as a candidate for open cloze questions. More details on this strategy can be found in (Pino, Heilman, and Eskenazi 2008).

Experimental Setup

The strategy described in the previous section produced acceptable open cloze questions 71% of the time. We applied it to 471 words in the Academic Word List (Coxhead 2000) and generated open cloze questions for these words. We selected 60 questions manually, keeping the ones that had the smallest number of possible answers. Questions for words of less than four letters were also discarded. An example is provided in Figure 1.

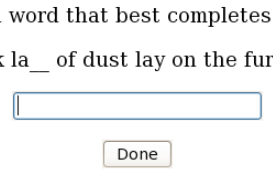

Figure 1: Example of open cloze question

Thirty one participants answered the selected questions in a session of approximately 45 minutes. 5 participants did not complete the task and their answers were discarded. Among the remaining 26 participants, 17 were non-native English speakers. 3 participants were from China and 14 from India. Thus participants were split into three groups: low proficiency speakers, high proficiency speakers and native speakers. The questions were randomly ordered and then given to the participants in the same random order. Therefore the

questions were not presented in the alphabetical order of the missing word; otherwise the order could have been used as an additional hint. The size of the blank was not an indication of the size of the correct answer. The questions were displayed alternatively in four different conditions: in condition 0 (C_0), the question had no hint; in Condition 1 (C_1), the first letter of the missing word was given; in Condition 2 (C_2), the first two letters were given; in Condition 3 (C_3), the first three. Six participants started in C_0, eight started in C_1, six started in C_2 and six started in C_3. For example, if a participant started in C_0, the first question would be displayed with no hint, the second question with the first letter of the missing words, etc. Thus each question was seen six times in C_0, eight times in C_1, six times in C_2 and six times in C_3. This accounts for the intrinsic difficulty of the questions regardless of the number of letters provided. An answer was considered to be correct when it was the expected word or the expected word with a misspelling (Laufer and Nation 1999).

Results and Discussion

In order to obtain a measure of the information provided by the letters, we investigated the following features:

- Number of letters provided
- Vowels or consonants provided
- Letter, bigram and trigram frequency
- Presence of syllable and syllable stress
- Syllable frequency

We also investigated the participant response time.

Number of Letters

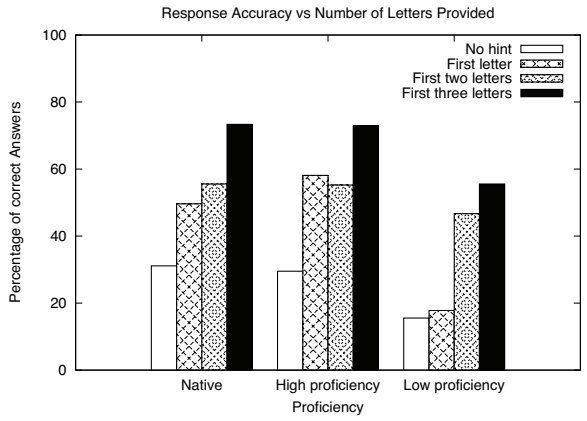

Figure 2: Proportion of correct answers vs. number of letters provided

Figure 2 shows the proportion of correct answers versus the number of letters provided. As expected, in general the proportion of correct answers increases with the number of letters provided. Indeed, providing the first letters of the words helps low proficiency speakers because it reminds them of a vocabulary word they might have forgotten and it helps high proficiency speakers who can often come up with something acceptable, but not the best possible answer because their vocabulary is richer. We have marked as acceptable answers that were not the expected ones but were synonyms or fit in the sentences. For C_0, there were fifty-eight distinct acceptable answers, for C_1 there were eighteen, for C_2 there were eleven and for C_3 there were eight. Most of the alternative answers are provided by native speakers and high proficiency speakers.

While low proficiency speakers got the highest improvement between C_1 and C_2, high proficiency and native speakers' performances confirmed the trend observed over all participants. High proficiency speakers underwent a decrease in performance between C_1 and C_2.

For native speakers and high proficiency speakers, the differences in performance between C_0 and C_1 and between C_2 and C_3 are statistically significant for a two-sided proportion test (native speakers C_0-C_1: $p = 0.003$; native speakers C_2-C_3: $p = 0.003$; high proficiency speakers C_0-C_1: $p =$ 6.5e-09; high proficiency speakers C_2-C_3: $p = 0.0003$). The difference between C_1 and C_2 is not statistically significant (native speakers: $p = 0.39$; high proficiency speakers: $p =$ 0.62). For low proficiency speakers, there were no significant differences between conditions. The significance tests and the shape of the graph suggest that providing the first letter versus the first three letters is critical to lowering question difficulty. Since there is no linear relation between the number of letters provided and the response accuracy, we need to further investigate the hint features mentioned above.

The participants spent 20.80 seconds on average per question for C_0, 22.98 seconds for C_1, 20.63 seconds for C_2 and 15.92 seconds for C_3. Thus the time spent on the questions does not directly reflect how effective the hints are: there is a significant increase in performance between conditions C_0 and C_1 but the time spent on the questions also increases between conditions 0 and 1 instead of decreasing as expected. On the other hand, the large decrease in time from C_2 to C_3 reflects the increase in performance between these two conditions.

Vowel versus Consonant

The previous section has underlined the role of the first and the third letter for word completion. Here we delve into this role and study the influence of these letters when they are vowels versus when they are consonants. Table 1 summarizes the results.

	Vowel	Consonant	p-value
Accuracy (1st letter)	45%	52.96%	0.18
Time (1st Letter)	23.96s	22.54s	0.62
Accuracy(3rd Letter)	81.11%	68%	0.01
Time (3rd Letter)	14.29s	16.40s	0.27

Table 1: Response Accuracy vs. First Letter and Third Letter as Vowels or Consonants

The role of the first and third letter is reversed: it seems that the first letter as a consonant helps the students more

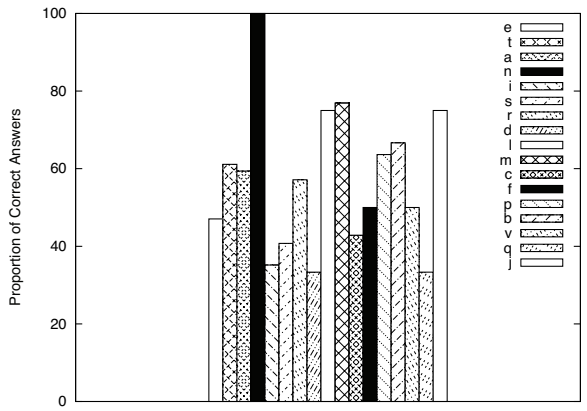

Figure 3: Performance vs. unigram frequency

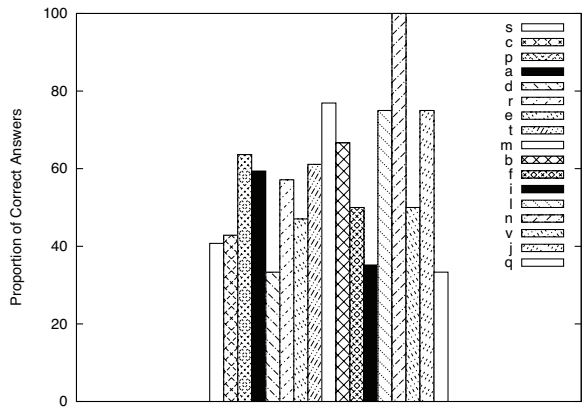

Figure 4: Performance vs. unigram *type* frequency

than the first letter as a vowel, while the third letter is more helpful when it is a vowel. We verified this trend within each linguistic group, American, Indian and Chinese. The response time also confirms this trend although response time differences are not significant. This result prevents us from drawing conclusions on the influence of individual letters as vowels or consonants. We need to turn to a another type of unit, namely letter n-grams and syllables, in order to discover relevant measures for the information provided by the letters.

Letter n-gram frequency

We evaluate the influence of letter n-gram frequency on question difficulty. Figures 3 and 4 show response accuracy in comparison with both *token* unigram frequency and *type* unigram frequency. The n-grams are ordered by decreasing frequency. Token frequencies were computed on a 9 MB corpus extracted from Project Gutenberg [1]. N-gram type frequencies were computed by counting the number of words that start with a given n-gram in a lexicon. We used Kilgarriff's lexicon (Kilgarriff 1997). Note that the type frequency order is different from the token frequency order. Performance does not seem to depend on the frequency of the provided unigrams. This also applies to bigrams and trigrams for which we did not draw graphs because of limited space. Analysis of response time shows that the average amount of time spent on each question does not depend on type or token frequency of the provided n-gram. Analysis of performance and response time within each linguistic group also confirms that n-gram frequency does not have an influence on question difficulty. Because of high variability in response accuracy, we also analyzed performances for the 25% most frequent n-grams, the next 25%, etc. This analysis did not show an influence of frequency on performance.

We conclude that n-grams are not the right unit to consider when measuring hint levels, therefore we further investigate other units, namely syllables.

[1] http://www.gutenberg.org

Syllables

Table 2 shows a comparison of students' performance depending on how many letters are provided and whether these letters form a whole syllable or part of a syllable. We determined which letters formed the first syllable using the CELEX database (Baayen, Piepenbrock, and Gulikers 1995). Note that the "Accuracy No Syllable" rows count the number of correct answers when the first n letters are provided and the first syllable has a length *greater* than n, not a length *different* from n. When the first letter or the first two letters were provided, students performed better when these two letters form a syllable. In that case, the difference is not significant; we believe that this is due to a small amount of data when the data is split by condition. When the first three letters were provided, the converse happened and the difference was statistically significant. However, the words that had a first syllable of length greater than three were usually short[2], which might explain this result. When comparing these two categories without regard of length, there is a significant increase in performance when a syllable is provided. Response time follows the same trend as response accuracy. Thus we conclude that rather than the number of letters provided, it is the presence of the first syllable that determines the question difficulty. We therefore rely on syllables to measure the amount of information provided by hint letters.

Since the presence or absence of syllables seems to play an important role in question difficulty, we further explored the influence of syllable frequency. We investigate both *token* and *type* frequency. CELEX database (Baayen, Piepenbrock, and Gulikers 1995) provides token frequency for a syllable in a given position or in any position. For example, the syllable "b{n" (phonetic transcription for "ban") has a frequency per million tokens of 31 in position 1, 64 in position 2, 0 in position 3 and 95 in any position. We did not find a relation between question difficulty and syllable token frequency. In order to compute type frequency, we count

[2] Such words were *couple*, *maintain*, *phase*, *quote*, *range*, *source*, *strategy* and *trace*

the number of words in the CELEX database that contain a syllable at a certain position or at any position. Students' performances did not appear to depend on syllable type frequency. Therefore we do not currently retain frequency as a measure of the information provided by the letters.

Number of Letters	1	2	3
Accuracy Overall	50.51%	54.36%	71.03%
Accuracy No Syllable	50.00%	50.00%	90.38%
Accuracy Syllable	66.67%	57.74%	68.45%
p-value (overall vs. syl.)	0.42	0.52	0.61
p-value (no syl. vs. syl.)	0.40	0.16	0.003
Accuracy No Syllable	53.28%		
Accuracy Syllable	63.22%		
p-value	0.003		
Time Overall	22.98s	20.63s	15.92s
Time No Syllable	23.27s	24.74s	11.73s
Time Syllable	13.58s	16.41s	15.18s
p-value (overall vs. syl.)	0.017	0.027	0.66
p-value (no syl. vs. syl.)	0.018	0.002	0.07
Time No Syllable	22.76s		
Time No Syllable	15.72s		
p-value	1.98 e-8		
American			
Accuracy Overall	49.63%	55.55%	73.33%
Accuracy No Syllable	48.82%	55.42%	92.86%
Accuracy Syllable	62.5%	54%	75.47%
Indian			
Accuracy Overall	58.10%	55.24%	72.86%
Accuracy No Syllable	57.77%	46.81%	91.43%
Accuracy Syllable	75%	60.38%	67%
Chinese			
Accuracy Overall	17.38%	46.67%	55.56%
Accuracy No Syllable	57.77%	46.81%	91.43%
Accuracy Syllable	75%	60.38%	67%

Table 2: Comparison between number of letters and syllable

Syllable Stress

After having shown the important role of syllables, we want to evaluate the role of syllable stress in question difficulty. Table 3 summarizes results for different types of stressed syllables. The first syllables of words were tagged either as "primary stress syllables", "secondary stress syllables" and "unstressed syllables" (Hayes 1995). Syllables with primary stress and no stress seem to give equivalent results; providing secondary stress syllables significantly decreases the students' performances. This applies when we consider syllables of a specific length or syllables with any length. The trend is confirmed when considering each population individually. We conclude that main stress syllables and non stressed syllables provide more information on the missing words than secondary stress syllables.

Number of Letters	1	2	3
Primary Stress (1)	-[a]	60.23%	71.15
Secondary Stress (2)	-	20%	55%
No Stress (0)	-	66.67%	68.18%
Primary Stress	66.15%		
Secondary Stress	37.5%		
No Stress	68.18%		
p-value (1 vs. 2)	5.49e-8		
p-value (1 vs. 0)	0.92		
p-value (2 vs. 0)	2.74e-7		
American			
Primary Stress	72.22%		
Secondary Stress	35.29%		
No Stress	67.5%		
Indian			
Primary Stress	65.04%		
Secondary Stress	39.13%		
No Stress	70.15%		
Chinese			
Primary Stress	53.33%		
Secondary Stress	-		
No Stress	44.44%		

[a] not enough data for this cell

Table 3: Performances for Different Syllable Types

Question Difficulty Hierarchical Levels

In the previous sections, we have explored different factors that influence question difficulty. We retain these factors as a measure of the information transmitted by the first letters of the words. We can use them in order to compare two questions with the heuristic exposed below. This algorithm returns the more difficult question given two questions `Question_1` and `Question_2` where `n_1` and `n_2` letters (designated by `l_1` and `l_2`) for the target words `tw_1` and `tw_2` are provided. The predicates `syl(x)` and `2str(x)` mean that `x` is respectively a syllable and that `x` is a secondary stress syllable.

```
Input:
Question_1
(tw_1, n_1),
Question_2
(tw_2, n_2)
if (n_1 < n_2)
  then return Question_1
else if (n_1 > n_2)
  then return Question_2
else if (syl(l_2) and not syl(l_1))
  then return Question_1
else if (syl(l_1) and not syl(l_2))
  then return Question_2
else if (2str(l_1) and not 2str(l_2))
  then return Question_1
else if (2str(l_2) and not 2str(l_1))
  then return Question_2
else more difficult question unknown
```

It should be noted that this algorithm is designed to work with open cloze questions that have the same amount of context. A sentence with poor context can be difficult to answer even if many letters of the missing word are provided.

Conclusion

In this study, we attempted to define a measure of the information transmitted by hints in open cloze questions when hints consist of the first few letters of the correct answer. We examined how the set of letters provided can convey different pieces of information and studied the influence of each piece of information on response accuracy. Our analysis showed that it is important to consider different aspects of lexical units such as letters, letter properties (vowel, consonant), and syllables. More specifically, we demonstrated that an important factor in determining question difficulty is not only the number of letters provided but also how these letters are assembled into syllables. Measuring the information transmitted by hints allows to control the amount of help provided by hint letters and to vary the difficulty level of questions in order to adapt it to student levels.

In this paper, we have focused on open cloze questions. However, it is also possible to use the first few letters of a word as a hint in a multiple choice question provided that all choices have the same first letter. Since multiple choice questions are widely used in many disciplines as a practice and assessment tool, the technique presented here can be adapted to other domains than English as a Second Language.

Acknowledgments

We would like to thank Sharon Rosenfeld for reviewing this paper. This research is supported by NSF grant SBE-0354420. Any opinions, findings, conclusions or recommendations expressed in this material are the authors', and do not necessarily reflect those of the sponsor.

References

Baayen, R.; Piepenbrock, R.; and Gulikers, L. 1995. The CELEX lexical database (CD-ROM). *Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA*.

Carreiras, M., and Perea, M. 2004. Naming pseudowords in Spanish: Effects of syllable frequency. *Brain and Language* 90(1-3):393–400.

Collins-Thompson, K., and Callan, J. 2005. Predicting Reading Difficulty With Statistical Language Models. *Journal of the American Society for Information Science and Technology* 56(13):1448–1462.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly* 34(2):213–238.

Cutler, A.; Mehler, J.; Norris, D.; and Segui, J. 1986. The syllable's differing role in the segmentation of French and English. *Journal of memory and language* 25(4):385–400.

Cutler, A. 1997. The Syllable's Role in the Segmentation of Stress Languages. *Language and Cognitive Processes* 12(5-6):839–846.

Hayes, B. 1995. *Metrical Stress Theory: Principles and Case Studies*. University Of Chicago Press.

Heilman, M.; Collins-Thompson, K.; Callan, J.; and Eskenazi, M. 2006. Classroom success of an intelligent tutoring system for lexical practice and reading comprehension. In *Proceedings of the Ninth International Conference on Spoken Language Processing.*

Kilgarriff, A. 1997. Putting frequencies in the dictionary. *International Journal of Lexicography* 10(2):135–155.

Klein, D., and Manning, C. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 423–430. Association for Computational Linguistics Morristown, NJ, USA.

Klein-Braley, C., and Raatz, U. 1984. A survey of research on the C-test. *Language Testing* 1(2):134–146.

Laufer, B., and Nation, P. 1999. A vocabulary-size test of controlled productive ability. *Language Testing* 16(1):33–51.

Lima, S. D., and Inhoff, A. W. 1985. Lexical Access During Eye Fixations in Reading: Effects of Word-Initial Letter Sequence. *Journal of Experimental Psychology* 11(3):272–285.

Nation, P. 2001. *Learning vocabulary in another language*. Cambridge University Press New York.

Oléron, P., and Danset, A. 1963. Données sur l'appréhension des mots. *Psychologie Française* 8:28–35.

Peretz, I.; Lussier, I.; and Béland, R. 1998. The Differential Role of Syllabic Structure in Stem Completion for French and English. *European Journal of Cognitive Psychology* 10(1):75–112.

Pino, J.; Heilman, M.; and Eskenazi, M. 2008. A selection strategy to improve cloze question quality. In *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems.*

Wood, H., and Wood, D. 1999. Help seeking, learning and contingent tutoring. *Computers & Education* 33(2-3):153–169.