Grammaticality Judgement in a Word Completion Task

Alfred Renaud[2] and **Fraser Shein**[1,2] and **Vivian Tsang**[1]

[1]Bloorview Kids Rehab
150 Kilgour Road
Toronto, ON M4G 1R8, Canada

[2]Quillsoft Ltd.
2416 Queen Street East
Toronto, ON M2A 1N1, Canada

{arenaud,fshein,vtsang}@bloorview.ca

Abstract

In this paper, we present findings from a human judgement task we conducted on the effectiveness of syntax filtering in a word completion task. Human participants were asked to review a series of incomplete sentences and identify which words from accompanying lists extend the expressions in a grammatically appropriate way. The accompanying word lists were generated by two word completion systems (our own plus a third-party commercial system) where the ungrammatical items were filtered out. Overall, participants agreed more, to a statistically significant degree, with the syntax-filtered systems than with baseline. However, further analysis suggests that syntax filtering alone does not necessarily improve the overall acceptability and usability of the word completion output. Given that word completion is typically employed in applications to aid writing, unlike other NLP tasks, accounting for the role of writer vs. reader becomes critical. Evaluating word completion and, more generally, applications for alternative and augmentative communication (AAC) will be discussed.

1 Introduction

Writers often need help from others to help with spelling and grammar. For persons with physical or learning disabilities, writing can be very stressful because of a greater reliance on the assistance of others. Software tools such as word completion are now commonly used to reduce the physical and cognitive load of completing a word or a sentence and thereby reducing a writer's dependence on others. But can such tools be as effective as a human with adequate linguistic knowledge? While it is hardly possible to completely emulate a human tutor or a communication partner, the purpose of this research is to investigate how much linguistic knowledge is necessary to ensure the usability of word completion. Here, we will focus on the grammaticality of word completion.

1.1 Word Completion

Word completion facilitates text entry by suggesting a list of words that can follow a given linguistic context. If the desired word is in the list, the user can select that word with a mouse click or a keystroke, thereby saving the effort of typing the remaining letters of the word. Otherwise, the user can continue typing while the software continues to display new lists of words based on that input.

For example, consider a user wants to type "That girl by the benches…" After each letter the user manually enters, a system would return a list of potential next words. Say, the next letter the user enters is "w." A system may offer the following choices: a) was, b) were, c) with, d) where, e) wrapped. By suggesting words in any given context, word completion can assist in the composition of well-formed text.

Typical word completion systems suggest words by exploiting *n*-gram Markov statistical models (Bentrup, 1987). These systems probabilistically determine the current word in a sentence given the previous *n*–1 words as context, based on a pre-generated *n*-gram language model derived from a corpus. With *n* typically being of low or-

Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing, pages 15–23,
Los Angeles, California, June 2010.

der (two or three, due to sparse data and computational issues), one consequence is that the applicability of suggested words beyond a certain word distance may become somewhat arbitrary. Further design improvements for word completion depend on the user population and the intended use. For example, the demand on the system to have a sophisticated language model may depend on whether the intent is to primarily reduce the physical or cognitive load of entering text. Evaluation approaches can elucidate on design and implementation issues for providing meaningful word choices.

1.2 Evaluation Approaches

A number of studies have been carried out to evaluate the efficacy of word completion systems. Koester (1994) measured *time savings*, which is the reduction in time that the user takes to generate a particular text with the aid of a word completion system compared to the time taken without it. The rationale for this measure is that any word completion system imposes a cognitive load on its users, whereby they now need to 1) change their focus between the target document and the word list display, and possibly between the screen and keyboard; 2) visually scan the word list to decide whether their intended word is present; and 3) select the intended word with the keyboard or mouse. Others have also examined similar visual-cognitive issues of using word completion (e.g., Tam and Wells, 2009). The overall approach implicitly defines a user-centred approach to evaluation by having human subjects simulate the actual writing process (usually in a copying, not writing task). Thus, results depend on the abilities and preferences of individual subjects.

System-based evaluation measures exist, the most common of which is *keystroke savings*. This measures the reduction in the number of keystrokes needed to produce a given text with the aid of a word completion system. Keystroke savings is an important factor for users with physical disabilities who have difficulty working with a keyboard for which it is desirable to keep the number of keystrokes to a minimum. A complementary measure, *completion keystrokes*, determines how quickly a given word is predicted by counting the number of characters required to reach completion. Completion keystrokes differs from keystroke savings in that the latter counts the letters remaining in the word.

In contrast to the previous two measures, both of which measure at the character level, *hit rate* measures at the word level by calculating the ratio of the number of words correctly predicted to the total number of words predicted. Given a sufficiently large lexicon, hit rate can be as high as 100% if every letter of every word is manually entered to its completion. As this can be misleading, hit rate is more typically measured with reference to the number of characters already typed in order to assess the system's demand on the user.

These objective measures address motor load independent of cognitive load. With the exception of time savings, these measures can be benchmarked automatically by simulating the writing process by using existing texts.

A shortcoming of these objective measures is that they focus on the reduction on the user's physical demand by simulating the entering of an already written text, and effectively ignore consideration of word choices other than the unique intended word. In reality, the actual writing process depends also on the quality of the entire group of suggested word choices with respect to the intended content. Renaud (2002) addressed this shortcoming by arguing that the syntactic and semantic relations between words can impact on choice-making at the target word. He introduced two measures, *validity* and *appropriateness*, measuring grammatical consistency and semantic relevance of *all* system output, respectively. The former measure calculates the proportion of a system's suggested words that is syntactically acceptable. The latter focuses on the proportion of relevant output based on lexical and domain semantics. Renaud compared a number of commercial systems and found a positive correlation between the new and existing measures. This finding also lends additional support to Wood's (1996) finding that offering syntactically and semantically appropriate choices improves performance. (Note that the converse may not hold true.)

For the remainder of this paper, we will put our emphasis on the impact of linguistic content (here, grammaticality) on the quality of word completion. The paper is organized as follows. In the next section, we will describe the need to incorporate syntactic information in word completion. In sections 3 and 4, we will describe our human judgement task evaluating the grammaticality of word completion. Based on our analysis, we will return to the evaluation issue in section 5 and discuss how grammaticality alone does not address the larger usability issue of word completion. Here, we propose that the word completion task, unlike traditional NLP tasks, requires both the reader's and writer's perspectives, which impacts the interpretation of our evaluation, and in turn impacts design decisions. In section 6, we will conclude by offering a more inclusive perspective on AAC.

2 The Demand for Syntactic Filtering

As shown earlier, many evaluation methods have focused on 1) the proportion of key-presses normally required during a typing session that the user need not to manually enter and 2) the proportion of text words in a typing session that the system is able correctly to predict. For a user with a learning disability or language difficulties, a greater concern is that all presented words be valid, logical, and free of grammatical errors. Current state-of-the-art systems suffer by suggesting words that are often syntactically implausible while excluding more justifiable but less probable suggestions (cf. our example in section 1). A user may be confused by inappropriate suggestions, even if correct suggestions are also present.

To quantify the importance of syntax in word completion, we compare the average hit rate scores (over all words) with the hit rate scores at points in sentences we consider as syntactically critical (see section 3 for their selection). Nantais et al. (2001) reported an overall hit rate of approximately 56% using bigram word completion after entering the first letter of a word across a large document. However, at the word location where it is crucial to maintain correct syntactic relation with the existing sentence fragment, hit rates are often much lower. In our study situation, the hit rate is at best 39%—these syntactic challenges tend to be semantically contentful and thus present difficulties to human subjects. Likewise, the systems are expected to struggle with them. Without a clear understanding of content specific issues during writing, examining time and keystroke savings alone does not reveal the increased difficulty a user faces at these word positions. We will return to these issues in section 5.

2.1 Building Syntactic Knowledge

Knowledge of syntax can be obtained by first tagging each dictionary word with its part of speech, such as noun or adjective. This information may then be used in either a probabilistic or a symbolic manner. Systems may reason probabilistically by combining tag *n*-gram models, where the part-of-speech tags for the previous *n*–1 words in a sentence are used to predict the tag for the current word, with word *n*-gram models that cue the resulting part(s) of speech to find words proper (Hunnicutt and Carlberger, 2001). Fazly and Hirst (2003) introduced two algorithms for combining tag trigrams with word bigrams. The first algorithm involved conditional independence assumptions between word and tag models, and the second algorithm involved a weighted linear combination of the two models.

A fundamental limitation to this approach is that low-order probabilistic language models can only account for relationships between closely co-located words. Symbolic syntactic prediction guided by a grammar, on the other hand, can deal with long-distance word relationships of arbitrary depth by applying rules that govern how words from syntactic categories can be joined, to assign all sentence words to a category. This approach uses knowledge of English grammar to analyze the structure of the sentence in progress and determine the applicable syntactic categories (e.g., noun, verb), along with other features (e.g., singular, past participle), to which the currently typed/predicted word must belong. In this way a word completion system is able to suggest words that are grammatically consistent with the active sentence fragment.

As such, research closer in nature to our work involves parsers that process the input sentence incrementally as each word is entered. Wood's (1996) augmented phrase-structure grammar showed that symbolic syntactic prediction can improve overall performance when combined with statistical orthographic prediction. McCoy (1998) used the augmented transition network or ATN (Woods, 1970) formalism to find candidate word categories from which to generate word lists. Gustavii and Pettersson (2003) used a chart parser to re-rank, or filter, word lists by grammatical value. These parsing algorithms manipulate some data structure that represents, and imposes ordering on, syntactic constituents of sentences. Recently, we have been developing a syntax module (Renaud et al., 2010) based on an ATN-style parser, which can facilitate both increasing the level of correctness in parses through grammar correction, and modifying the information collected during parsing for a particular application (Newman, 2007). Specifically, this system filters words provided by *n*-gram completion such that the word list only shows words that fit an acceptable grammatical structure. It operates on a longer list of the same frequency-ranked words our core predictor generates. Under this setup, our syntax module can influence the final list shown to the user by demoting implausible words that otherwise would have been displayed and replacing them with plausible words that otherwise would not. Our rationale for using a symbolic vs. a probabilistic parser in word completion is beyond the scope of the current paper.

3 Grammaticality Judgement Experiment

To evaluate the impact of syntactic filtering on word completion, we devised a human judgment task where human subjects were asked to judge the grammatical acceptability of a word offered by word completion software, with or without syntactic filtering. Given a partial sentence and a leading prefix for the next word, word completion software presents a number of choices for the potential next word. Although the goal is to assess the grammaticality of predicted words with or without syntactic filtering, the intent is to assess whether the inclusion of syntactic heuristics in the word completion algorithm improves the quality of word choices.

3.1 Experimental Setup

In our experiment, we compared three different word completion systems: our baseline completion system (WordQ®[*], henceforth "baseline"), our word completion system with syntax filtering ("System B"). We also included a third-party commercial word completion system with syntax filtering built-in (Co:Writer®[†], "System C"). In each system, we inputted a partial sentence plus the leading character for the next word. Each system returned a list of five choices for the potential next word. Our subjects were asked to judge the grammatical acceptability of each word (binary decision: yes or no).

It is worth noting that the more letters are manually inserted, the narrower the search space becomes for the next word. Nantais et al. (2001) suggested that after inserting two characters, the hit rate via automatic means can be as high as 72%; the hit rate for humans is likely much higher. Given that our goal is to examine the grammaticality of word choices and not hit rate, providing only one leading letter allows sufficient ambiguity on what the potential next word is, which in turn allows for a range of grammatical choices for our judgement task.

3.2 Sentence Selection

We selected our test sentences from Canadian news sources (Toronto Star and the Globe and Mail), which are considered reliably grammatical. We chose a total of 138 sentences.[‡] Each sentence was truncated into a fragment containing the first x-1 words and the first character of the x^{th} word, where x ranges from three to ten inclusive. The truncation position x was deliberately selected to include a variety of grammatical challenges.

We divided the sentence fragments into nine types of grammatical challenges: 1) subject-verb agreement; 2) subject-verb agreement in question-asking; 3) subject-verb agreement within a relative clause; 4) appositives; 5) verb sequence (auxiliary verb-main verb agreement); 6) case agreement; 7) non-finite clauses; 8) number agreement; and 9) others.

For example, the sentence "That girl by the benches was in my high school" from section 1.1 can be used to test the system's ability to recognize subject-verb agreement if we truncate the sentence to produce the fragment "That girl by the benches w___." Here, subject-verb agreement should be decided against the subject "girl" and not the (tempting) subject "benches."

[*] http://www.wordq.com; our baseline system uses a bigram language model trained on a corpus of well-edited text.

[†] http://www.donjohnston.com/products/cowriter/index.html

[‡] We did not pick a larger number of sentences due to the time constraint in our experimental setup. The rationale is to avoid over-fatiguing our human subjects (approximately an hour per session). Based on our pilot study, we were able to fit 140 sentences over three one-hour sessions.

After the initial selection process, we reduced our collection to 123 partial sentences. Because the sentences were not evenly distributed across the nine categories, we divided the sentences into three sets such that the frequency distribution of the sentence types was the same for all three sets (41 sentences per set). The three word completion systems were each assigned a different set.[§]

3.3 Grammaticality Judgements

We fed each partial sentence into the corresponding system to produce a word list for grammatical judgement. Recall our example earlier, given five word choices per partial sentence, for each word choice, our subjects were asked to judge its grammatical acceptability (yes or no).

We recruited 14 human subjects, all native speakers of English with a university education. Each subject was presented all 123 sentences covering the three systems, in a paper-based task. The sentence order was randomized and the subjects were unaware of which system produced what list.

Given that each system produced a list of five options for each partial sentence, each subject produced 5×41=205 judgements for each system. There were 14 sets of such judgements in total.

4 Results and Analysis

Our primary objective is to examine the subjects' agreement with the system, and whether the subjects generally agree among themselves. Our rationale is this. If the subjects generally agree with one another, then there is an overall agreement on the perception of grammaticality in word completion. If this is indeed the case, we then need to examine how and why our subjects agree or disagree with the systems. Otherwise, if there is low inter-subject agreement, aside from issues related to the experimental setup, we need to reconsider whether offering grammatical word completion choices is indeed practical and possible.

We first calculated individual participant agreement with the output of each system (i.e., averaged over all participants). The baseline scored 68%. System B scored 72% and System C scored 74%. Thus, an early important result was that syntax assistance in general, independent of particular approach or algorithm, does appear to improve subject agreement in a word completion task. (Note that we treat system C as a black box as we are not privy to its algorithms, which are not published.)

Overall, the grammaticality of a given test word (i.e., averaged over all test words) had an average agreement of 85%, or by 12 of the 14 participants. The percentage agreement for each system was 84% for the baseline, 87% for system B, and 86% for system C. If at least two-thirds of the participants (10 of 14) agreed on the grammaticality of a particular test word, we considered the collective opinion to be consistent for that word and declared a consensus. Participants reached consensus on 77% of the test words for the baseline, 82% of the test words for system B, and 80% of the test words for system C.

Next, we calculated consensus participant agreement for each system. This measure was different from the previous in that we considered only those cases where 10 or more of the 14 participants agreed with one another on the grammaticality of a system's test word and discarded all other cases. In 75% of the consensus cases for the baseline, the subjects agreed with the system (by approving on the grammaticality); in the other 25% of the consensus cases the subjects disagreed with the system. System B scored 78% on the consensus agreement and system C scored 81%.

A repeated-measures analysis of variance (ANOVA) was performed on the data. For both individual and consensus participant agreement, each of Systems B and C outperformed the baseline system (statistically significant, p<.05), while the difference between the two systems with syntax awareness was not statistically significant.

To summarize our findings, our subjects generally found the output grammatically more acceptable if syntactic assistance was built in (72% and 74% over 68% in raw participant agreement; 78% and 81% over 75% in consensus agreement). The behaviour of our System B generally was in line with the behaviour of the third-party System C. Finally, the agreement among subjects for all systems was quite high (~85%) and is considered reliable.

[§] We initially to used three different sets, i.e., one set per system, to avoid a sampling "fluke" of different grammatical difficulties/categories. However, for exactly the same reason, we also tested our system using the two sets for the other two systems for ease of comparison. See section 4.1 for details.

4.1 Subject Agreement with Other Systems

To further understand the behaviour of our own system (in contrast to our subjects' judgements), we create two new systems, A' and C' based on the output of the baseline system and the third-party System C. Recall that the sentence set used in each system is mutually exclusive from the set used in another system. Therefore, this setup introduces an additional set of 41 sentences × 5 predicted words × 2 systems = 410 judgements.

Our setup is simple: we feed into our parser each of the sentence fragments for the corresponding system, along with each predicted word originally produced. If our parser accepts the word, the analysis remains unchanged. Otherwise, we count it as a "negative" result, which we explain below.

Consider again our earlier example, "The girl by the benches w___." Say system C' produces the following options: a) was, b) were, c) with, d) where, e) wrapped. We then attempt to generate a partial parse using the partial sentence with each predicted word, i.e., "The girl by the benches was," "The girl by the benches were," and so on. If, for instance, our parser could not generate a parse for "The girl by the benches where," then we would treat the word choice "where" as not approved for the purpose of recalculating subject agreement. So if any subjects had approved its grammaticality (i.e., considered it a grammatical next word), then we counted it as a disagreement (between the parser and the human judge), otherwise, we considered it an agreement.

Consider the following example. One partial sentence for System C was "Japanese farmers immediately pick the shoots which a[m]..." Only 1 of 14 judges agreed with it. System C' also flagged "am" as ungrammatical. Now 13 judges agreed with it.

On the other hand, consider this partial sentence originally from the baseline system, "The reason we are doing these i[nclude]..." where 10 judges said yes but our parser could not generate a parse. In this case, A' scores 4 on agreement.

Overall, A' overrode 10 decisions and scored 71% agreement as a result. That is a 3% improvement over the baseline 68% score. Nine of the 10 reversed consensus in a positive direction and 1 (example above) reversed consensus in a negative direction. In comparison, C' overrode 6 decisions, and scored 76% (2.0% improvement over the original 74%). Five of 6 cases reversed consensus, all in a positive direction. (The other case reversed a non-consensus in a positive direction.) Given that the theoretical maximum agreements for the two systems are 84% and 86% (i.e., regardless of polarity), there is considerable increase in the subject agreement.

It is worth noting that many subjects made the number agreement mistake due to proximity. In the previous example, "The reason we are doing these i[nclude]...", the subjects made the incorrect agreement linking "include" to "these" instead of linking to "the reason." While these cases are not prevalent, this is one reason (among many) that the theoretical maximum agreement is not 100%.

4.2 System's vs. Subjects' Perspective

Although the agreement between the systems and the subjects were high, no system achieved perfect agreement—many words were considered ungrammatical extensions of the partial sentences. We see two possible explanations: 1) the disagreeable output was erroneous; or 2) the disagreeable output was grammatical but judged as ungrammatical under certain conditions.

We manually examined the parse trees of the "disagreeable" cases from our system. Interestingly, in most cases, we found there exists a reasonable parse tree leading to a grammatical sentence. We thus conclude that grammaticality judgements of partial sentences might not completely reflect the underlying improvement of the word completion quality. That is, discrepancies between human and computer judgement need not point to a poor quality syntax filter; instead, it may indicate that the system is exhibiting correct behaviour but simply disagrees with subjects on the particular grammatical cases in question. In such cases, subjects' disagreement with the system does not provide sufficient grounds for making modifications to the system's behaviour. Rather, it is worth examining the factors leading to the subjects' perception of a word as an ungrammatical extension of a partial sentence.

5 Discussion

Overall, our results indicate that our subjects agree with the grammaticality of word completion more when syntactic filtering is used than not.

That said, in light of the disagreeable cases, we believe that the quality of word completion may not be so straightforwardly evaluated.

5.1 Selectional Restriction

Take this example, “The plane carrying the soldiers a___.” The next word “are” was unanimously considered ungrammatical by our human judges. Consider the following full sentence version of it: “The plane carrying the soldiers are contemplating is too difficult a task.” In this case, the subject is “the plane carrying” (as an activity), the relative clause is “the soldiers are contemplating”, and finally, the verb phrase is “is too difficult a task.” This sentence may be difficult to interpret but a meaningful interpretation is possible syntactically and semantically.

Consider the following variation, “The political situation the soldiers a___.” In this case, it is not difficult to conceive that “are” is a possible next word, as in “The political situation the soldiers are discussing is getting worse.” The syntactic construction is [noun phrase] [relative clause] [verb phrase]. Both partial sentences have a potential grammatical parse. Why then is one considered grammatical and the other not?

Sentences that induce midpoint reading difficulties in humans are well known in psycholinguistics and are referred to as garden-path sentences (Frazier, 1978). Reading “the plane carrying the soldiers” induces an expectation in the reader’s mind that the sentence is about the plane doing the carrying, and not about the carrying of the plane by the soldiers, leading to a “short circuit” at the word “are.”

In linguistics and CL, one aspect of this phenomenon, selectional restriction, has been explored previously (most notably Levin, 1993 and Resnik, 1995). Selectional restriction is defined as the semantics of a verb restricting the type of words and phrases that can occur as its arguments. Essentially, the meaning of the verb makes an impact on what is possible syntactically and semantically. What we observe here is a generalized case where it is no longer only about a verb placing syntactic and semantic restrictions on its surrounding words. Instead, we observe how a word or a number of words influencing the semantic interpretation, and in turn impacting on the perception of grammaticality of the next word (cf. hit rate issues in section 2).

5.2 Evaluation Approach

Although our original intent was to study the grammaticality of word completion, ultimately the question is what impacts on the quality of word completion. It is without a doubt that the grammaticality of the next word suggestions impacts on the perception of the quality of word completion. However, we believe the key hinges on whose perspective of quality is considered, which then becomes a usability issue.

Recall that word completion is designed to aid the writing process. The curious part of our evaluation was that we devised it as a grammaticality judgement task via reading. Is grammaticality different when one is reading vs. writing? We consider this issue in two ways.

Partial Sentences vs. Full Sentences

Let us revisit our garden-path example:

1a. The plane carrying the soldiers a[re]…
1b. The plane carrying the soldiers are contemplating is not that difficult a task.
2a. The political situation the soldiers a[re]…
2b. The political situation the soldiers are losing sleep over is getting worse.

In sentences 1a and 2a, readers have no choice but to judge the grammaticality of “are” based on the existing partial sentence. Depending on the reader’s creativity, one may or may not anticipate potential full sentences such as 1b and 2b. In contrast, consider an alternative experimental setup where the readers were offered full sentences such as 1b and 2b and were asked to judge the grammaticality of “are.” Given the complexity of the sentences (selectional restriction aside), the readers would have no choice but to consider the existence of a relative clause, which should increase the likelihood of evaluating “are” as a grammatical component of the sentence.

Reading vs. Writing

Now we have observed the potential impact on grammaticality judgements of a potential next word when reading a partial sentence vs. a full sentence. That said, it needs emphasizing that the key issue is to evaluate the quality of a suggested next word given a partial sentence, not grammaticality in complete isolation. When a user uses word completion, he/she is actively engaged in the writing process. No software can truly predict the intent of the writer; the full sentence is waiting to be written and cannot be written a priori.

Consider someone who is in the process of writing the sentence "The plane carrying the soldiers..." Is this writer likely to be debating in his/her head whether the sentence is about the plane that does the carrying or "plane carrying" as an activity? Clearly, the writer's intent is clear to the writer him/herself. In contrast, a sentence may be perfectly grammatical and semantically reasonable, yet a reader may still find it ambiguous and/or difficult to read. In other words, the perception of grammaticality of a next word depends on the task (reading vs. writing). This is *not* to say that our evaluation task is compromised as a result. Despite that the general grammar rules do not change, our reading judgements depending on the context (e.g., partial vs. full sentence) suggests that the reading perspective only provide a partial picture on the quality of output that is intended for a writing task. In our case, higher quality syntactic filtering (e.g., our parser here) may not lead to greater usability.

6 Concluding Remarks

In this paper, we have shown that the quality of word completions depends on the perspective one takes. Considering that AAC is to aid someone in producing content for communication, i.e., for third-party consumption, the reading-writing dichotomy is too serious an issue to ignore. This issue has received some CL attention (Morris, 2004, 2010; Hirst, 2008, 2009) but has not been discussed in the AAC literature (Tsang et al., 2010). The question remains, how do we then evaluate, and more generally, design and use an AAC application?

We believe the issue is far from clear. Take our current focus—grammaticality of word completion. If the form of the content produced is ungrammatical or difficult to read from the perspective of a reader, you risk having the reader misunderstand the writer's intent. However, from the writer's perspective, unless he/she is perceptive of the interpretation problems with his/her potential readers, there is no incentive to produce content as such; the writer can only produce content based on his/her previous linguistic experience.

One may argue that corpus statistics may best capture human linguistic behaviour. For example, hit rate statistics using existing corpora is one such way of assessing the quality of word completion. However, corpora tell only one half of the story—only the writing half is captured, the interpretation issues from the reading side are rarely captured, if at all.

More important, the design of word completion is setup in a way that the task consists of both a reading component and a writing one—the appropriateness of suggested words is assessed by the writer via reading during the writing task. In fact, this is not merely a case of reading vs. writing, but rather, an issue of relevance depending on the linguistic context as well as the user's perception of it. Traditionally, researchers in CL and psycholinguistics have attempted to deal with human processing of linguistic content at various levels (cf. the CUNY Conference on Human Sentence Processing, e.g., Merlo and Stevenson, 2002). However, no computational means is truly privy to the content behind the linguistic form. Content, ultimately, resides in the reader's or the writer's head, i.e., intent. The question remains how best to design AAC to aid someone to communicate this content.

In summary, in our grammaticality judgement task, incorporating syntax in word completion improves the perceived quality of word choices. That said, it is unclear how quality relates to usability. Indeed, the evaluation is far from conclusive in that it only captures the reader's perspective and not the writer's. Currently, we are not aware of the existence of a purely writer-based evaluation for grammaticality of word completion (see Lesher et al., 2002 for one curious attempt). More generally, the reader-writer (or speaker-listener) dichotomy is unexplored in AAC research and should be considered more seriously because communication (as text, speech, or otherwise) involves multiple people producing and consuming content, where the perception of content differs considerably. The challenge of AAC may lie in bridging the gap between production and consumption where communication is neither only about communicating intent nor making interpretations.

Acknowledgments

This project is funded by Quillsoft Ltd. We also wish to thank Jiafei Niu (University of Toronto) for conducting our usability study and Frank Rudzicz (University of Toronto) for providing helpful comments.

References

John Bentrup. 1987. Exploiting Word Frequencies and their Sequential Dependencies. *Proceedings of RESNA 10th Annual Conference*, 121–122.

Afsaneh Fazly and Graeme Hirst. 2003. Testing the Efficacy of Part-of-Speech Information in Word Completion. *Proceedings of the 2003 EACL Workshop on Language Modeling for Text Entry Methods*, 9–16.

Lyn Frazier. 1978. *On Comprehending Sentences: Syntactic Parsing Strategies*. Ph.D. Thesis, University of Connecticut.

Ebba Gustavii and Eva Pettersson. 2003. *A Swedish Grammar for Word Prediction*. Master's Thesis, Department of Linguistics, Uppsala University.

Graeme Hirst. 2008. The Future of Text-Meaning in Computational Linguistics. In *Proceedings of the 11th International Conference on Text, Speech and Dialogue*, 1–9.

Graeme Hirst. 2009. Limitations of the Philosophy of Language Understanding Implicit in Computational Linguistics. In *Proceedings of the Seventh European Conference on Computing and Philosophy*, 108–109.

Sheri Hunnicutt and Johan Carlberger. 2001. Improving Word Prediction Using Markov Models and Heuristic Methods. *Augmentative and Alternative Communication*, 17(4):255–264.

Heidi Koester. 1994. *User Performance with Augmentative Communication Systems: Measurements and Models*. Ph.D. thesis, University of Michigan.

Gregory W. Lesher, Bryan J. Moulton, D. Jeffery Higginbotham, and Brenna Alsofrom. 2002. Limits of Human Word Prediction Performance. In *Proceedings of 2002 CSUN Conference*.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Kathleen F. McCoy. 1998. *The Intelligent Word Prediction Project*. University of Delaware. http://ww.asel.udel.edu/nli/nlp/wpredict.html

Paola Merlo and Suzanne Stevenson, Eds. 2002. *The Lexical Basis of Sentence Processing: Formal, Computational and Experimental Issues*. John Benjamins Publishing Company.

Jane Morris. 2004. Readers' Interpretations of Lexical Cohesion in Text. *Conference of the Canadian Association for Information Science*, Winnipeg, Manitoba.

Jane Morris. 2010. Individual Differences in the Interpretation of Text: Implications for Information Science. *Journal of the American Society for Information Science and Technology*, 61(1):141–149.

Tom Nantais, Fraser Shein, and Mattias Johansson. 2001. Efficacy of the word prediction algorithm in WordQ. In *Proceedings of the 2001 RESNA Annual Conference*, 77–79.

Paula S. Newman. 2007. RH: A Retro Hybrid Parser. In *Proceedings of the 2007 NAACL Conference*, Companion, 121–124.

Alfred Renaud. 2002. *Diagnostic Evaluation Measures for Improving Performance of Word Prediction Systems*. Master's Thesis, School of Computer Science, University of Waterloo.

Alfred Renaud, Fraser Shein, and Vivian Tsang. 2010. A Symbolic Approach to Parsing in the Context of Word Completion. In Preparation.

Philip Resnik. 1995. Selectional Constraints: An Information-Theoretic Model and its Computational Realization. *Cognition*, 61:127–125.

Cynthia Tam and David Wells. 2009. Evaluating the Benefits of Displaying Word Prediction Lists on a Personal Digital Assistant at the Keyboard Level. *Assistive Technology*, 21:105–114.

Vivian Tsang and Kelvin Leung. 2010. An Ecological Perspective of Communication With or Without AAC Use. In Preparation.

Matthew Wood. 1996. *Syntactic Pre-Processing in Single-Word Prediction for Disabled People*. Ph.D. Thesis, Department of Computer Science, University of Bristol.

William Woods. 1970. Transition Network Grammars for Natural Language Analysis. *Communications of the ACM*, 13(10):591–606.