Corpus-based Analysis of Modals in Consecutive Sentences

Robert Chartrand*, Hidenobu Kunichika**, and Akira Takeuchi*

*Department of Artificial Intelligence, **Department of Creation Informatics,
Faculty of Computer Science, Kyushu Institute of Technology
680-4 Kawazu, Iizuka-shi, Fukuoka, 820-8502, Japan
robert@minnie.ai.kyutech.ac.jp, kunitika@minnie.ai.kyutech.ac.jp, takeuchi@minnie.ai.kyutech.ac.jp

This work presents a method that was developed for extracting examples that include modal auxiliaries in two consecutive sentences from an English corpus. The method uses a process where the sentences are extracted, parsed for simplification of complex sentences, analyzed for frequency of use and then made available for display to provide useful examples for the purpose of learning the use of modals in context. These example sentences can be useful tools for both teachers and students alike.

1. Introduction

Modal auxiliaries are among the most difficult structures to teach to students of English as a second language (Celce-Murcia & Larsen-Freeman, 1999). Other languages often use different structures to convey the ideas expressed in English by modals, learners frequently make mistakes with modals (Coelho, 2004). Although there are a large number of grammar books that explain the form and meaning of the English modals, it is difficult to understand the meaning of modals merely by explanation.

One solution is to learn the meaning/usage by applying corpus linguistics methods (Meyer, 2002) and show learners examples of authentic usage of modals. Research on the use of modality in the English language by using corpus linguistics has been reported by performing a simple concordance of modals for the benefit of learners using the World Wide Web for research on the use of epistemic modality and the Michigan Corpus of Academic Spoken English (MICASE) was used to research the use of deontic modality (Meyer, 2006).

In this paper, we aim to demonstrate a method to extract useful examples from two consecutive sentences using modals. We conducted our analysis by using the British National Corpus (BNC) due to the availability and affordability of obtaining a broad collection of English samples from a wide range of sources (Oxford, 2008).

2. Consecutive use of modals

In order to understand why two consecutive sentences using modals would be useful for learners of English, let us explore the usage of modals. Modals have both semantic and formal properties, which do not readily make themselves understood with simple chunks of the language. Language teachers often have great difficulty in teaching the inherent complexity of modals. Thus, capturing this complexity by using different examples so that the learners can internalize the meaning into their minds remains an important objective. One common problem with modals, for example, occurs with the words, “would and should.” If the following modals were given without the sentences to provide context, the learner would have more difficulty in understanding the situation. Look at Example 1:

IT REALLY WOULD WORK AFTER ALL
WE SHOULD KNOW SOON ENOUGH

Example 1. Consecutive extracted modals

The combination of the two sentences makes the flow of thought more intuitive and offers a greater chance of comprehension to the learner by using deductive reasoning. The use of context and vocabulary allow for the learner to relate these two modals consecutively and appropriately. A single sentence will not have as much intrinsic impact and will offer no suggestions as to how to employ another modal in the vicinity of the first modal.

3. Method of extracting example sentences

While the BNC contains a large number of useful sentences containing modals, the process of automatically extracting modals from sentences included in the BNC involves certain difficulties.
A simple string match was not sufficient for our purposes, therefore it was necessary to construct an elaborate method for extracting sentences from the corpus.

First, only the main clauses in sentences were taken into account. Some sentences in the BNC are complex sentences. Therefore, there could be a possibility of multiple modal auxiliaries occurring, and thus only the main clause of each sentence was selected and the subordinate clauses were deleted from the data. Moreover, we determined that by removing the subordinate clauses, this made the connotation of the sentence easier to understand for English language learners, and when presenting two consecutive sentences, the meaning becomes more evident. There is also a possibility that a single sentence from the BNC includes multiple main clauses. In this case, all main clauses are extracted as separate sentences. To achieve this task, it was necessary to parse the sentences automatically, thus we used the Charniak Parser (Charniak, 2000), (Charniak & Johnson, 2005). After parsing the sentences, we processed the output, i.e. parse trees, in order to recognize which parts were the main and the subordinate clauses.

Second, when extracting consecutive sentences containing modals, both paragraph boundaries and text source boundaries were taken into account. Because we wanted to extract a consecutive use of modals from an equivalent situation, the two sentences should be contained in the same paragraph of a given text in the corpus.

Third, after the sentences were modified to remove the unnecessary subordinate clauses, the t-scores were calculated to determine the most likely occurrence of the modals. In corpus linguistics, the t-score refers to how many instances of the co-occurring word are found in a designated span between of the node word (the first modal) and the co-occurring word (the second modal), and how many instances might be expected in that span, given the frequency of the co-occurring word in the corpus as a whole. In addition, the t-score uses a calculation of standard deviation which takes into account the probability of co-occurrence of the node and its collate and the number of tokens in the designated span in all lines. Therefore, it maybe assumed that high t-score results in some non-random association between the two modals, whereas low t-score results indicate low frequency use. Here, we have defined an equation to calculate the t-score for modals in consecutive sentences, as shown in equation (1), where:

\[ t-score = \left( f(n,c) - (f(n) \times f(c)/N) \right) \sqrt{f(n,c)} \]  

\( f(n) \) refers to the frequency of the sentences containing the node modal,

\( f(c) \) refers to the frequency of the sentences containing the collate modal,

\( f(n,c) \) refers to the frequency of the sentences containing the collate modal,

\( N \) refers to the number of sentences in the corpus.

From the result of extracting the main clauses by parsing, we extracted modals in consecutive sentences and calculated the t-scores. We can thus use this information for selecting practical example sentences.

4. T-score results

In this study, we analyzed the BNC for occurrence of modals in two consecutive sentences. We identify the first occurring modal (node word) as Modal 1 (M1), and the modal occurring in the second sentence (collate) as Modal 2 (M2). The results are shown in Table 1. The modals that were investigated are listed on the left side (M1), as well as on the top of the table (M2). The t-scores are listed in the corresponding cells. It can be seen, for example, that the pair combination “can-will” has a t-score of 9.4, which is a relatively high number. Similarly, the combination “may-should” has a t-score of 7.7, which is also high. Therefore, it can be said that these two pairs of combinations are respectively strongly associated modals that are found in the BNC, and correspondingly, occurring frequently in English. To verify the authenticity of these results, let’s look at Example 2 from the data extracted from the BNC

T-score = 9.4
\( f(\text{can}) = 97686, f(\text{will}) = 160434, f(\text{can,will}) = 3427, n = 5423057 \)

CAN YOU GIVE ME ONE MORE DAY TO COME UP WITH SOMETHING

I WILL SLEEP ON IT TONIGHT AND TOMORROW

Example 2 High occurrence example of “can, will”

As Table 1 demonstrates, some pair combinations of modals are more common than others, and as the t-score number is higher, so is
the prospect of finding two modals in consecutive sentences that are prevalent. The high values along the diagonal represent the consecutive uses of the same modal in both sentences, such as “will, will” with a t-score of 110.5. The negative t-scores represent the modal combinations that are less frequent, it does not mean, however, that they are bad examples of English use.

5. Discussion

Learning by which concordances of specific language features can be analyzed by learners to infer and test their meanings, can be made up of two forms: inductive and deductive learning (Aston, 2001). The inductive form encourages learners to infer meaning from the output data by identifying patterns and making up some generalizations by observation. The deductive form encourages learners to apply previously acquired knowledge to classify the data according to the grammar rules they have already learned. Moreover, this type of data-driven learning (DDL) may involve both induction and deduction, since arriving at an understanding of the raw data may require not only inducing the information but also testing it deductively to verify that it makes sense (Murison-Bowie, 1996). Thus a learner may read useful examples of modals and induce the meaning, and consecutively use deductive reasoning of grammar rules to verify that the meaning is correct.

As noted earlier in the Introduction, modal auxiliaries are difficult to teach, and one of the most effective ways to learn how to use modals for English learners is to study examples of how they are used in practice. It is more efficient to teach modals by placing them in context. The best way to show this is to give examples (Harris & McLaughlin, 1997). Due to the nature and the complexity of modals, rules simply cannot be memorized and applied for general cases (Celce-Murcia & Larsen-Freeman, 1999). Thus, a holistic approach to learning modals, by observing examples of modals in a specific situation, may be better than by memorizing rules of use. A holistic approach by concordancing can be developed by assembling a set of language data to form the basis of teaching and learning for EFL students, as learners need to engage actively in processing the meanings of the language (Willis, 2000). Recent studies in corpus linguistics have shown that having students access various corpora and drawing their attention to the concordance evidence provides them with a more objective view of the language (Hirata & Hirata, 2007). Moreover, students pay attention to the target language in context to gain an extensive knowledge of the vocabulary in more effective ways (Cobb, 1997).

6. Conclusion

When considering materials for teaching or learning, one of the most valuable tools is having definite concrete examples in context that facilitate the learner’s ability to process the language. It has been shown here that by using specific methods for retrieval of sentences in an English language corpus, along with parsing and analysis, one can obtain a large amount of useful data for learning purposes. This data can be used by a learner in a number of situations. Due to the large amount of data, a search can be conducted to find contextual information and useful examples can be retrieved for inclusion of learning tasks. The t-score information available in Table 1 can be used as a guide to teach the more useful combinations of modals.

Moreover, a learner could also study the use of modals by accessing a database with the

<table>
<thead>
<tr>
<th>M1/M2</th>
<th>can</th>
<th>must</th>
<th>will</th>
<th>may</th>
<th>shall</th>
<th>should</th>
<th>could</th>
<th>would</th>
<th>might</th>
</tr>
</thead>
<tbody>
<tr>
<td>can</td>
<td>56</td>
<td>6.1</td>
<td>9.4</td>
<td>15.9</td>
<td>1.4</td>
<td>0.9</td>
<td>-22.8</td>
<td>-40</td>
<td>-1.4</td>
</tr>
<tr>
<td>must</td>
<td>6.6</td>
<td>36</td>
<td>-0.5</td>
<td>3.1</td>
<td>1.9</td>
<td>5.9</td>
<td>-9.3</td>
<td>-9.9</td>
<td>-5.1</td>
</tr>
<tr>
<td>will</td>
<td>11.8</td>
<td>-7.8</td>
<td>110.5</td>
<td>10.8</td>
<td>7.8</td>
<td>2.2</td>
<td>-41.8</td>
<td>-68.5</td>
<td>-13.4</td>
</tr>
<tr>
<td>may</td>
<td>11.3</td>
<td>3.4</td>
<td>12.9</td>
<td>62.2</td>
<td>2.9</td>
<td>7.7</td>
<td>-11.1</td>
<td>-22.8</td>
<td>3.1</td>
</tr>
<tr>
<td>shall</td>
<td>-0.6</td>
<td>1.4</td>
<td>7.9</td>
<td>3.1</td>
<td>25.2</td>
<td>-1.2</td>
<td>-19.1</td>
<td>-16.7</td>
<td>-2.8</td>
</tr>
<tr>
<td>should</td>
<td>3.8</td>
<td>5.4</td>
<td>1.1</td>
<td>5.8</td>
<td>-1.2</td>
<td>48.3</td>
<td>-3.2</td>
<td>0.4</td>
<td>2.3</td>
</tr>
<tr>
<td>could</td>
<td>-31.6</td>
<td>-9.7</td>
<td>-50.5</td>
<td>-24.2</td>
<td>-12.1</td>
<td>-9.8</td>
<td>37.6</td>
<td>25.8</td>
<td>11</td>
</tr>
<tr>
<td>would</td>
<td>-45.5</td>
<td>-14.2</td>
<td>-66.3</td>
<td>-37.1</td>
<td>-12.5</td>
<td>-10.7</td>
<td>22.3</td>
<td>95.1</td>
<td>14.1</td>
</tr>
<tr>
<td>might</td>
<td>-6.9</td>
<td>-6.8</td>
<td>-16.1</td>
<td>-1.9</td>
<td>-1.4</td>
<td>-0.3</td>
<td>10.4</td>
<td>17.1</td>
<td>24.6</td>
</tr>
</tbody>
</table>

Table 1. T-scores of modals in consecutive sentences
information presented in this paper, in order to become more familiar with expressions that are specific to the learner’s needs. Such a user could identify certain verbs and degrees of complexity that would fit particular criteria. Diverse levels of users could have access to distinctive parts of a database. A student for example, would have different needs than a teacher, and thus would approach the use of the database in a different way. The use of the information presented here in a database is a work in progress and accessibility through the Internet of this database is an objective of further study and development.

The retrieval of information from a corpus and the processing of the result is important to find practical examples of modals in consecutive sentences. Therefore, a method of dealing with vast amounts of information is necessary to achieve this purpose. In summary our process can be described as follows:

- We identified a corpus to be the source of our text files. In this case we used the BNC due to its large size and XML functionality.
- We eliminated the spoken portion of the corpus because modals are not often used in concurrent sentences in natural speech.
- We ascertained that the consecutive sentences originated from the same paragraph of a story to establish the continuation of an idea.
- Parsing of sentences was performed.
- The main clause of each sentence was kept and the subordinate clause was deleted. This simplified the sentences for easy comprehension.
- We extracted sentences that contained modal auxiliaries in two consecutive sentences.
- The t-scores were calculated to determine the most likely occurrence of the modals.
- The t-scores were examined and useful examples were extracted from the data.

One of the problems with this method is the use of parsing software to break down complex sentences (Heift & Schulze, 2007) and identify the main clause and subordinate clause. The Charniak parser is well known, however, it is not flawless and there are some inaccurate results that could be output from the program (Charniak, 2000). Complex passive sentences are especially difficult to parse accurately and this could affect the subordinate clause determination.

Ultimately, it is our objective to develop a web-based solution to incorporate all aspects of this method to allow a user to set search parameters in order to produce an output that would be useful as a learning tool. Such parameters could be the choice of the modals, the level of complexity, and length of the example sentences. Learners and educators could make use of these results to improve the quality of writing, a better understanding of modals and improving the process of learning a difficult aspect of the English language.

References


Willis, J. A holistic approach to task-based course design. The Language Teacher, 24(2), 7-11 (2000).