Analysis Units in Listening Recall: Comparison Between Idea Unit, Pausal Unit, Tone Unit, and Proposition

Hideki IIMURA
Tokyo University

Abstract

This study aimed to explore the analysis units that would be suitable for listening recall. It compared four analysis units—idea unit, pausal unit, tone unit, and proposition. Forty-seven participants were asked to listen to a passage twice and write down everything that they could remember from it. The recall protocols were scored against the four analysis units. Their score on each of the four analysis units was compared to an external criterion measure, the listening section of the computerized assessment system for English communication (CASEC). The results indicated that all the analysis units were significantly correlated with the external measure. Further, this study analyzed the data using factor and cluster analyses. The results of these analyses revealed the characteristics of each analysis unit. Finally, this study identified the tone unit to be the most suitable analysis unit for listening recall.

1. Introduction

A recall task (sometimes referred to as a free written recall or an immediate recall) is a widely used method of assessing L2 listening comprehension. According to Sakai (2005), the recall task should be a valid and effective L2 listening comprehension measure that reveals the learners’ listening process. In this task, students are asked to listen to a passage and write down everything they can remember from it. Students’ responses—that is, recall protocols—are scored according to a list from the original passage and divided into analysis units. Hence, the selection of an analysis unit is an essential decision in the assessment of recall protocols.

However, researchers thus far have been confronted with the task of deciding the analysis unit that should be adopted for listening recall. As shown in Table 1, there is a variety of analysis units being used at the present time; of this variety, some analysis units have been defined in different ways (idea unit and proposition), some have not been defined (meaning chunk), and some have been
interpreted as another analysis unit (proposition as information unit*). Further, there has been little comparative research concerning analysis units for listening recall, which have been considered as options that researchers must select from.

<table>
<thead>
<tr>
<th>Analysis units</th>
<th>Previous studies using analysis units</th>
<th>Definitions of analysis units</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Markham &amp; Latham (1987)</td>
<td>Spencer’s system</td>
</tr>
<tr>
<td></td>
<td>VanPatten (1990), Vogely (1995)</td>
<td></td>
</tr>
<tr>
<td>Information unit</td>
<td>Dunkel &amp; Davis (1994), Jung (2003b)</td>
<td>Turner &amp; Greene (1977)</td>
</tr>
</tbody>
</table>

Based on an experimental study, this study represents an attempt to determine the analysis units from among idea unit (IU), pausal unit (PU), tone unit (TU), and proposition (PRO), which are appropriate for listening recall. Therefore, this study will compare them in terms of validity with an external criterion and will also reveal their characteristics using a cluster analysis.

2. Methods

2.1 Participants

This study involved 47 Japanese university students. They were all second-year students majoring in English.

2.2 Materials

2.2.1 Proficiency test

This study is an attempt to investigate the analysis units from among the four that will yield higher concurrent validity. The computerized assessment system for English communication (CASEC), developed by the Japan Institute for Educational Measurement, Inc., was used as the external criterion. This test system measures English proficiency based on the item response theory. It consists of the following four sections: (1) vocabulary, (2) phrasing, (3) listening, and (4) dictation. This study used the listening section of the CASEC since it aims to explore the relationship between recall performances and the listening proficiency level. The CASEC was found to have high reliability ranging from .964 to .975 (Hayashi, 2001).
2.2.2 Listening recall test

A 62-word passage, tape-recorded by a female native English speaker, was chosen from the listening section of Part 3 in the STEP test for the pre-second grade held in October 2003 (Obunsha, 2005). The participants were instructed to listen to the passage twice and write down everything that they remembered. They were required to produce their recall protocols in Japanese “so as not to interfere with their ability to demonstrate comprehension” (Heinz, 2004, p. 102). They were not allowed to take notes while listening. The time for writing recall protocols was approximately three minutes.

2.3 Analyses

2.3.1 Definitions of analysis units and the experimental passage

This study used four analysis units—IU, PU, TU, and PRO—for scoring recall protocols. The researcher and another coder independently divided the passage based on the definitions of each analysis unit; following this, discrepancies were discussed until both the coders reached an agreement. The following was the rate of agreement between them: IU, 78.6%; PU, 58.3% (66.7%); TU, 81.0%; and PRO, 79.3%. At PU, the percentage provided in parentheses shows the rate of agreement over assigning importance to each item of the text. As stated above, the IU and PRO have been defined in several ways. This study employed the definitions provided by Carrell (1985) and Turner and Green (1977), respectively. The decision to use both definitions was made because in their articles, they included samples of dividing the text. The following are the definitions of the four analysis units, and the passage was segmented using each analysis unit.

Idea unit analysis based on Carrell (1985)

Each idea unit consisted of a single clause (main or subordinate, including adverbial and relative clauses). Each infinitival construction, gerundive, nominalized verb phrase, and conjunct was also identified as a separate idea unit. In addition, optional and/or heavy prepositional phrases were also designated as separate idea units. (p.737)

1.On Saturday Robert went out for dinner. / 2.On his way home, / 3.he found a woman’s handbag. / 4.He looked inside / 5.for a name / 6.or telephone number, / 7.but he couldn’t find any information about the owner, / 8.so he took the bag / 9.to the police station. / 10.Two days later, / 11.a police officer called / 12.to say / 13.that the owner had come / 14.to pick up her bag.

Pausal unit analysis based on Johnson (1970)

The function of a pause might be to catch a breath, to give emphasis to the story, or to enhance meaning. The raters then calibrated the units according to their structural importance by assigning
each a value between one and four — one (1) being the least and four (4) being the highest. In other words, the passage was divided into “pause acceptable units,” twenty-five percent of which were given a significance value of (1); twenty-five percent (2), twenty-five percent (3), and twenty-five percent (4). (Reported in Schmidt-Rinehart, 1994, p.187)

1. On Saturday / 2. Robert went out for dinner / 3. On his way home / 4. He found a woman’s handbag / 5. He looked inside / 6. For a name or telephone number / 7. But he couldn’t find any information about the owner / 8. So he took the bag to the police station / 9. Two days later / 10. A police officer called to say / 11. That the owner had come / 12. To pick up her bag.

Value
Level 1: [1, 3, 9], Level 2: [2, 6, 10], Level 3: [5, 7, 12], and Level 4: [4, 8, 11]

Tone unit analysis based on Brazil (1994)

Each tone unit of ordinary speech has either one or two prominent syllables. The last prominent syllable in each tone unit is also a tonic syllable. The tonic syllable is the place at which a significant pitch movement or tone begins. (p.8)


Proposition based on Turner & Greene (1977)

Propositions consist of abstract word concepts. ...The propositions themselves consist of two or more word concepts, such that the first is a relation and the others arguments. The relation is a connection between a set of arguments forming a single idea. Propositions have been classified into three distinct classes, based on the type of relation they contain: Predication, Modification, and Connection2. (p.2)

1. (GO, A: ROBERT, G: OUT)
2. (GET, A: ROBERT, O: DINNER)
3. (PURPOSE: FOR, 1 > 2)
4. (TIME: ON, P1, A2: SATURDAY)
5. (POSSESS, A: WOMAN, O: HANDBAG)
6. (FIND, A: ROBERT, O: P5)
7. (GO, A: ROBERT, G: HOME)
8. (TIME: WHEN, P6, P7)
9. (DISJUNCTION: OR, NAME, TELEPHONE NUMBER)
10. (LOOK, A: ROBERT, G: INSIDE)
>11. (GET, A: ROBERT, O: P9)
12. (PURPOSE: FOR, P10, >11)
13. (EXTENT OF: INFORMATION, ANY)
14. (QUALITY OF: INFORMATION, OWNER)
2.3.2 Scoring

Recall protocols were scored against each analysis unit of the original text. In order to ensure the reliability of scoring, the researcher and a second rater scored all the data, with an interrater reliability coefficient of .92. Discrepancies were discussed until the two raters reached an agreement.

3. Results & Discussion

Table 2 shows the descriptive statistics of the listening recall test. The following analyses were performed using the percent correct scores in order to make easier comparisons among all the analysis units.

Table 2  Descriptive Statistics of the Listening Recall Test

<table>
<thead>
<tr>
<th></th>
<th>N = 47</th>
<th>M (%)</th>
<th>SD (%)</th>
<th>The number of items</th>
<th>Full mark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idea unit</td>
<td>6.9 (49.5)</td>
<td>6.9 (29.6)</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Pausal unit</td>
<td>13.2 (44.0)</td>
<td>10.2 (34.2)</td>
<td>12</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Tone unit</td>
<td>11.7 (55.9)</td>
<td>6.0 (28.8)</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Proposition</td>
<td>10.0 (34.5)</td>
<td>9.3 (32.6)</td>
<td>29</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

The results of ANOVA with repeated measures reveal that there was a significant difference between the four analysis units ($F (1.8, 81.2) = 52.914, p < .001$). To further investigate this difference, a Bonferroni test was performed. This post hoc test reveals that the TU was the highest, followed by the IU, the PU, and then the PRO ($p < .01$) (tone unit > idea unit > pausal unit > proposition).

This order might be related to the number of items in the passage divided by the four analysis units. In other words, the more items a passage has, the more marks the participants obtain. Take the example of the first sentence in the passage: “On Saturday Robert went out for dinner.” This sentence was divided into as many as 4 items using the TU analysis, where the participants obtained
only a mark if they merely wrote down “Robert.” On the other hand, with both the IU and PU, the participants could not obtain a mark by writing down merely “Robert.” If this was the case, why were the scores of the PRO lowest of the four analysis units? This can be explained by the distinctive characteristic of PRO’s scoring system. As described in 2.3.2, [ROBERT] appears not only in P.1 but also in P.2, P.3, P.4, P.6, P.7, and others. A total of 14 items include [ROBERT] in the passage. When protocols are scored using the PRO, if “Robert” is not recalled, not only P.1 but also the remaining 13 items are automatically regarded as incorrect responses. Accordingly, it can be said that the PRO is the strictest among the four analysis units, whereas the TU is the most lenient.

For the purpose of determining how well the four analysis units relate to an outside criterion, Pearson product-moment correlations between the percent correct scores of the four analysis units and the listening section of CASEC (CASECL) were calculated. As shown in Table 3, all the four analysis units and CASECL were significantly correlated; the highest correlation was observed with regard to the TU, followed by the IU, the PRO, and the PU.

| Table 3 Pearson Correlations Between the Four Analysis Units and CASECL |
|---------------------------------|---------|---------|---------|---------|---------|
|                                  | IU      | PU      | TU      | PRO     | CASECL  |
| Idea unit                        |         | .978**  | .973**  | .892**  | .408**  |
| Pausal unit                      | -       | .980**  |         | .885**  | .398**  |
| Tone unit                        | -       |         | .890**  |         | .431**  |
| Proposition                      | -       |         |         | .400**  |         |
| CASECL                           |         |         |         |         | -       |

**p < .01

Moreover, Table 4 presents high correlations among the four analysis units ranging from .885 (between the PU and the PRO) to .980 (between the PU and the TU). A factor analysis was conducted in order to examine the relationship among the four analysis units. As expected, one factor was extracted, which accounted for 93.4% of the total variance. Table 4 reveals the results of factor analysis. These convergent results, together with significant correlations between the analysis units and CASECL, provide positive evidence of validity for the external aspects of all the analysis units used in this study.

<table>
<thead>
<tr>
<th>Table 4 Results of Factor Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Component matrix*</td>
</tr>
<tr>
<td>Idea unit</td>
</tr>
<tr>
<td>Pausal unit</td>
</tr>
<tr>
<td>Tone unit</td>
</tr>
<tr>
<td>Proposition</td>
</tr>
</tbody>
</table>

*Maximum likelihood method

This study assumed that recall protocols would have some distinct pattern and that analyzing
these patterns could provide us the key to reveal the characteristics of the analysis units. Thus, a cluster analysis using the Ward method was employed. The four clusters were identified (Figure 1).

**Figure 1  Dendrogram of Cluster Analysis Using the Ward Method**

The means and standard deviations of the percent correct scores by the four clusters are presented in Table 5 and illustrated in Figure 2.

**Table 5  Means and Standard Deviations of the Percent Correct Scores by the Four Clusters**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>IU</th>
<th>PU</th>
<th>TU</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>M(%)</td>
<td>SD</td>
<td>M(%)</td>
<td>SD</td>
</tr>
<tr>
<td>1 (16)</td>
<td>34.4</td>
<td>10.2</td>
<td>25.4</td>
<td>14.0</td>
</tr>
<tr>
<td>2 (11)</td>
<td>16.9</td>
<td>4.8</td>
<td>6.7</td>
<td>2.6</td>
</tr>
<tr>
<td>3 (10)</td>
<td>92.1</td>
<td>10.4</td>
<td>92.0</td>
<td>9.3</td>
</tr>
<tr>
<td>4 (10)</td>
<td>67.1</td>
<td>9.6</td>
<td>66.7</td>
<td>11.1</td>
</tr>
</tbody>
</table>

**Figure 2  Mean Profile of the Four Clusters**
Four ANOVAs were conducted to assess whether the mean percentages of the four clusters varied across the four analysis units. The results indicated that there were significant differences across the four clusters: IU ($F(3, 43) = 146.53, p < .01$); PU ($F(3, 43) = 142.42, p < .01$); TU ($F(3, 43) = 206.58, p < .01$), and PRO ($F(3, 43) = 138.37, p < .01$). Tukey's HSD post hoc comparisons were made between the four clusters on each of the four analysis units. The cluster means of the percent correct scores are presented in Table 6 for each of the analysis unit. In the table, these means are arranged in a descending order (from left to right). The clusters that were not statistically different from one another are enclosed in parentheses.

<table>
<thead>
<tr>
<th>Table 6</th>
<th>Means of the Percent Correct Scores by Cluster and Post Hoc Comparisons*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 3 &gt; Cluster 4 &gt; Cluster 1 &gt; Cluster 2</td>
<td></td>
</tr>
<tr>
<td>IU</td>
<td>Pausal unit</td>
</tr>
<tr>
<td>92.0</td>
<td>94.8</td>
</tr>
<tr>
<td>34.1</td>
<td>25.4</td>
</tr>
<tr>
<td>16.9</td>
<td>22.5</td>
</tr>
</tbody>
</table>

* The > symbol indicates statistical significance, $p < .01$.

As stated above, the four clusters were identified with regard to the patterns of recall protocols. In Cluster 1, elevations and dips were observed on the IU and TU and the PU and PRO, respectively.

Cluster 2 demonstrated the same graphic pattern as Cluster 1: elevations on the IU and TU and dips on the PU and PRO. Based on these results, it is reasonable to say that for listeners who perform poorly on listening recall, both the PU and PRO might result in a boundary effect such that their scores are clustered toward or at the bottom of the distribution, whereas the IU and TU could identify even smaller clues to what the listeners comprehend.

Cluster 3 demonstrated a relatively flat profile with no remarkable patterns of dips or elevations across the four analysis units. This cluster was the highest in terms of elevation (i.e., listening proficiency). Based on this result, it can be interpreted that any analysis unit could be valid for listeners performing well in listening recall; perhaps, it would be more correct to say that there would be no difference among them.

Although the rate of correct performance for the participants in Cluster 4 was above or near the average rate, a sharp dip was observed on the PRO. This demographic pattern might be caused by the strict scoring system of the PRO mentioned above. In other words, the participants in this cluster may have failed to comprehend the repeated items in the text.

Considering the results and discussion above, the IU and TU can be regarded as suitable analysis units for listening recall since both are relatively highly correlated with the external criterion ($r = .408$ and $r = .438$, respectively) and can discriminate between the levels of listeners' performances at a wider range. The noteworthy point here is the difference between the IU and TU.
The IU has been widely used in reading research (Riley and Lee, 1996), and hence, it can be stated that this analysis unit is tailored in the written language. Conversely, as Limura (2003) noted, the TU is an analysis unit that takes into consideration certain acoustic effects on listening comprehension. Therefore, there is a possibility that the TU might be a more suitable analysis unit for listening recall.

4. Conclusion

The present study analyzed four analysis units—IU, PU, TU, and PRO—from various angles in relation to the following points: (1) the criterion-related validity of each of the four analysis units, (2) the interrelationship among the four analysis units using factor analysis, and (3) the characteristics of each of the four analysis units by grouping recall protocols.

The findings of this study indicated that the TU might be a more suitable analysis unit than the others for listening recall from the viewpoint of concurrent validity and the ability of different levels of listeners’ performances. In addition, as discussed in the preceding section, the superiority of the TU lies in its nature characterized by phonological features.

The researcher acknowledged the fact that the number of samples and tests carried out in this study was small. In order to confirm the findings of this study, further research should include increased number of participants and tests. In addition, it should include the practical aspect of analysis units since analyzing a text based on the definition of an analysis unit is a laborious task.

Acknowledgements

This research was supported in part by a Grant-in-Aid for Young Scientists (B) (17720138) from 2005 to 2007. I would like to thank Dr. Rie Koizumi for her valuable comments.

Notes

*1. Both Dunkel & Davis (1994) and Jung (2003b) regarded their analysis unit as the information unit suggested by Anderson (1980). However, Anderson (ibid) defined his analysis unit as proposition. Therefore, this study judged the information unit of their analysis unit to be proposition.

*2. For further details of this analysis unit division, see Turner and Green (1977, pp.5–27), where they describe the manner in which a text can be propositionized.

*3. The Greenhouse-Geisser corrected F value is used because the Mauchly Sphericity test is significant (p < .001).

References


