Vocabulary Learning and Instruction
Volume 9, Issue 2, August 2020
http://vli-journal.org

The Challenges of Measuring Multi-Word Expression Use in Conversation

Haidee Thomson
Hokusei Gakuen University Junior College Department and Victoria University of Wellington

Abstract

This article introduces three important challenges and possible solutions when using spoken dialogue to measure the use of specific multi-word expressions. The first challenge is deciding whether to count precise and accurate use of target expressions only or whether to extend the count to include variation. The second challenge requires addressing the indirect nature of dialogue as a testing method. The third challenge is organizing data and preparing ways to clearly identify speakers within the dialogue. These challenges are illustrated with examples and potential solutions from my recent research investigating spoken use of multi-word expressions.

1 Background

Learners of English in Japan are known to struggle with speaking fluency (Herder & Sholdt, 2014; Nishino & Watanabe, 2008). Research has suggested that speakers can boost their speaking fluency by using common multi-word expressions (Boers, Eyckmans, Kappel, Stengers, & Demecheleer, 2006; Wood, 2009). Therefore, replicable classroom interventions that are proven to increase knowledge and use of multi-word expressions would be very useful.

In response to the need for proven teaching interventions, I conducted a classroom-based study to measure participants' ability to use target multi-word expressions (such as *I think I will* and *how do I get*). This article does not report results but rather discusses the challenges and solutions found to measuring multi-word expression use. Multi-word expressions were sourced from classroom materials and checked whether they were in current general use by checking the occurrence of the expressions in the spoken sub-corpus of the Corpus of Contemporary American English (COCA) (Davies, 2013). There were 30 selected expressions, each made up of four words. Participants were from an engineering university in northern Japan, with minimal exposure to English outside of a 90-minute English class once a week (Thomson, 2017). This research was conducted as an action research project (Burns, 2010), wherein after each study was completed, results and feedback were used to inform adjustments for follow-up studies and replication. The experimental group was introduced to the target expressions through travel themed units where they were encouraged to use the expressions in activities such as shadowing, role-play, dictogloss, and more. The control group was not exposed to the target expressions, as they studied engineering topics in English using a

Please cite this article as: Thomson, H. (2020). The Challenges of Measuring Multi-word Expression Use in Conversation. *Vocabulary Learning and Instruction, 9* (2), 63–68. https://doi.org/10.7820/vli.v09.2.thomson

linked skills format. In order to assess the efficacy of the experimental teaching intervention, learners were randomly partnered together and given a scenario to role-play and audio record before and after the 6-week intervention. Participants then uploaded the audio recordings to a class Moodle page, where I could measure the use of the target expressions within the conversations.

As I endeavored to assess the impact of the classroom intervention through participant dialogue recordings, I faced three challenges. The first challenge was deciding whether to adhere strictly to counting the exact target expressions or whether to include variation when measuring multi-word expression use. This decision had the potential to limit or broaden the scope of conclusions about learning. The second challenge was the use of role-play, where one cannot be sure that unused expressions are actually unknown. Role-play allows speakers to choose their expressions, so that conclusions from such an indirect test are necessarily partial and methods of triangulation need to be considered. The third and final challenge regarded technical and logistical decisions about how to organize data collection and how to identify speakers. These decisions affected sample size and in turn how generalizable the results would be.

1.1 Challenge One: Decisions About How to Measure Multi-Word Expressions

When measuring multi-word expressions, it is important to be clear about what is to be counted and how it is to be counted, so that fair comparison or replication is possible (Porte, 2012). Running analysis on a small subset so that various counting options can be compared is likely to assist decision-making on the most appropriate counting method for a particular purpose and context. In order to identify whether the participants used the target expressions in my research, I first transcribed participant conversations (pre- and post-intervention) to Notepad UTF-8 files. I was then able to use AntConc (Anthony, 2014) to search for sequential words that made up the target expressions using the concordance word search function. Use of complete four-word expressions was rare, so in order to discover how much learning had actually taken place, a more sensitive measure was required. I decided to count partial use from two-word target combinations upward. There were also instances where participants used two or more words from a target expression with alternative words to complete the expression: for example, one target expression was "how do I get (to)," and a participant said, "how can I go (to)." The first and third words were from the target expression, whereas the second and fourth words were variants of the target. If the variant words are counted, then "how can I go" could be counted as four words. However, if only the target words are counted, then, "how can I go" could only be counted as two words. I chose to count both ways and found that including such instances of variation in the measurement did not change the overall comparison of results; it simply lifted the numbers. For instance, in my first study (of three), I investigated the length of multi-word expressions and found the overall comparison post-intervention between the control group (n = 8, Mdn = 2.35) and experimental group (n = 15, Mdn = 2.63) was not statistically different whether variation was counted U = 86.5, asymp p = 0.085, z = 1.72, r = 0.359 or not: control

group $Mdn = 2$, experimental group $Mdn = 2$, $U = 83.0$, asymp $p = 0.081$, $z = 1.75$, $r = 0.364$. There was a medium-sized effect between the experimental group and the control group either way. Therefore, for the sake of clarity, I decided to keep the measurement simple by only counting the target words used in target expressions in follow-up studies (not the variant words used in target expressions).

1.2 Challenge Two: Using Role-Play as an Indirect Test of Multi-Word Expression Knowledge

As my participants were a convenience sample from conversation-based classes, I wanted to use conversation in my tests. Measuring knowledge of specific multi-word expressions through use in natural-like conversation is an inherently indirect measurement method because speakers are under no compulsion to use the specific expressions being measured in conversation. Previous research that has used dialogue to measure language production includes Tavakoli (2016) who used a discussion task between two learners to compare fluency with a monologic task in an English as a Second Language (ESL) environment, while Taguchi (2007) used dialogue between learner and researcher to investigate the use of memorized chunks in Japanese as a foreign language. However, the use of dialogue between learners to investigate use of multi-word expressions seems to be rather unique. The beauty with scenario role-play is that learners are free to choose the words or multi-word expressions that they wish to use. If a speaker chooses to use an expression under the natural time pressure of conversation, it can be assumed to some degree that they know it. The flip-side of conversational freedom is that speakers can choose alternative expressions to complete the role-play. Participants may be able to retrieve and use the expressions being measured, but in the absence of task-essentialness they can complete the role-play without using them (Thomson, Boers, & Coxhead, 2019). There are a multitude of reasons why learners may choose not to use known vocabulary or expressions in their output, including nonnecessity or *lack of confidence* (see Coxhead, 2018).

While spoken and written modes are not the same, the use of target expressions in either speech or writing reflects some knowledge of the expression. My solution to the indirect nature of role-play was to directly test target expression knowledge through a cloze test after the role-play. In this way, knowledge of meaning and written form of the target expressions (as tested in the cloze test) could be triangulated with the ability to produce and use orally (as measured through the role-play). It was informative to be able to cross-check (triangulate) evidence of knowledge of expressions from the cloze test with use in the role-play. The combination of these two complementary measures revealed more direct evidence as to what extent the expressions were known.

1.3 Challenge Three: Technical and Logistical Challenges

When setting up a dialogue assessment situation, decisions need to be made about the allocation of partners, identification of individuals, and roles. It is preferable to keep the speaker partnerships the same for all tests. Every speaker has their own style of speaking in their first language (e.g., long pausing patterns or

quick-fire speaking), which may influence their second language speaking style (de Jong, Groenhout, Schoonen, & Hulstijn, 2015; Derwing, Munro, Thomson, & Rossiter, 2009). And, every combination of speakers is likely to have their own dynamic, as speaking patterns converge (de Jong, 2018; Pardo, 2013; Wilson & Wilson, 2005). I had a pre- and post-experiment model; so, in order to keep the conditions as similar as possible, I strived to have participants record their pre- and post-dialogue with the same partner, playing the same role, in the same scenario role-play. Absenteeism was difficult to control, and as a result there were different partnerships in some cases, which I had to exclude from my analysis. In class situations, absenteeism will impact continuity of conditions, so checking for schedule clashes and perhaps providing incentives may help to reduce sample shrinkage.

Identification of speakers was an important challenge to consider when designing the study. I chose to record conversations using audio only (rather than using video). I made this choice for two reasons: the availability of audio recording software that participants could operate simply, and also data capacity, as storing video data for many conversations would be problematic. In my first study, participant pairs labeled the audio file with their names but were not instructed to say their names or roles at the beginning of the recording. Unfortunately, when I listened to the recordings, I found that it was extremely difficult to match the pre-intervention voice with the post-intervention voice. Most of my participants were male, and many had similar voice tones. Therefore, I ran my speech analysis by dyad rather than individual, which halved my sample size (from 46 to 23). Naturally, halving the sample size reduced the power of the data set (Thomson, 2017). Learning from this experience, the next time I collected conversational data. I instructed the participants to say their names, identification numbers, and role at the beginning of the interaction, which helped me to associate their voice and name on each recording. For those considering dialogue analysis or assessment, the use of video recording might be worthwhile, as it would add visual evidence, which would help when identifying who is speaking.

In a before and after intervention measurement of conversation, ideally the speakers would play the same roles in both conversations. In a separate follow-up study, I instructed participants to play the same role as they had in the pre-test. However, in spite of this instruction, close to half of the group ended up playing opposite roles to what they had played in the pre-test. I excluded these dialogues, which reduced my sample size in the follow-up study from 52 to 26. A fairly simple initiative to ensure that participants play the same roles in pre- and post-dialogues would be to make a list of names with roles to be shared with participants prior to doing the final recording. Such planning is simple but can be easily overlooked.

2 Implications for Future Research

The three challenges described above for assessing knowledge and the ability to use multi-word expressions through dialogue can be mitigated through advanced consideration and planning. Firstly, decisions regarding which words to count and not to count need to be thoughtfully considered. Second, measuring target expression use from conversation can be fraught with difficulties because a speaker

may have been able to use an expression, but the opportunity to use it did not arise in the conversation or perhaps they chose to use an alternative word combination. Therefore, combining role-play with a more direct measure (such as a cloze test) can provide triangulation to help show the expressions that are known even if they are not used in the role-play. Finally, technical and logistical planning can maximize continuity between pre- and post-intervention conditions and dialogue partnerships. Anticipating and planning for these challenges will take some of the stress out of data analysis and/or assessment. I hope that these insights from my own data collection and analysis will help readers plan and implement their own spoken language assessments and research involving multi-word expressions.

References

Anthony, L. (2014). *AntConc (Version 3.4.3) [Computer Software]*. Tokyo, Japan: Waseda University. Retrieved from http://www.laurenceanthony.net/

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, *10*(3), 245–261. doi:10.1191/1362168806lr195oa

Burns, A. (2010). *Doing action research in English language teaching: A guide for practitioners*. New York: Routledge.

Coxhead, A. (2018). Vocabulary and second language writing. In J. I. Liontas, M. DelliCarpini, N. J. Anderson, D. D. Belcher, & A. Hirvela (Eds.), *The TESOL encyclopedia of English language teaching, first edition* (pp. 2597–2602). Hoboken, USA: John Wiley & Sons Inc. doi:10.1002/9781118784235.eelt0533

Davies, M. (2013). *Corpus of global web-based English: 1.9 billion words from speakers in 20 countries*. Retrieved from http://corpus.byu.edu/glowbe/

de Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237–254. doi:10.1080/15434303.2018.1477780

de Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*(2), 223–243. doi:10.1017/S0142716413000210

Derwing, T. M., Munro, M. J., Thomson, R. I., & Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(04), 533. doi:10.1017/S0272263109990015

Herder, S., & Sholdt, G. (2014). Employing a fluency-based approach to teach the TOEFL iBT: An action research project. In T. Muller, J. Adamson, P. S. Brown, & S. Herder (Eds.), *Exploring EFL fluency in Asia* (pp. 26–41). Hampshire, UK: Palgrave Macmillan.

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOL Quarterly*, *42*(1), 133–138. doi:10.2307/40264432

Pardo, J. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4,* 1–5. doi:10.3389/fpsyg.2013.00559

Porte, G. (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

Taguchi, N. (2007). Chunk learning and the development of spoken discourse in a Japanese as a foreign language classroom. *Language Teaching Research*, *11*(4), 433–457. doi:10.1177/1362168807080962

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *IRAL: International Review of Applied Linguistics in Language Teaching*, *54*(2), 133–150. doi:10.1515/iral-2016-9994

Thomson, H. (2017). Building speaking fluency with multiword expressions. *TESL Canada Journal*, *34*(3), 26–53. doi:10.18806/tesl.v34i3.1272

Thomson, H., Boers, F., & Coxhead, A. (2019). Replication research in pedagogical approaches to spoken fluency and formulaic sequences: A call for replication of Wood (2009) and Boers, Eyckmans, Kappel, Stengers & Demecheleer (2006). *Language Teaching*, *52*(3), 406–414. doi:10.1017/S0261444817000374

Wilson, M., & Wilson, T. P. (2005). An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, *12*(6), 957–968. doi:10.3758/BF03206432

Wood, D. (2009). Effects of focused instruction of formulaic sequences on fluent expression in second language narratives: A case study. *The Canadian Journal of Applied Linguistics*, *12*(1), 39–57.