

# From Dialogue to Learner Modeling: Identifying Candidate Signals of Productive Use in LLM-Based Grammar Practice

Luisa Ribeiro-Flucht<sup>1</sup>, Lanhua Huang<sup>2</sup>, Xiaobin Chen<sup>1</sup>,

<sup>1</sup>University of Tübingen Germany, <sup>2</sup>Guangzhou City University China,

[luisa.ribeiro-flucht@uni-tuebingen.de](mailto:luisa.ribeiro-flucht@uni-tuebingen.de)

[huanglh@gcu.edu.cn](mailto:huanglh@gcu.edu.cn)

[xiaobin.chen@uni-tuebingen.de](mailto:xiaobin.chen@uni-tuebingen.de)

## Abstract

Adaptive language-learning systems often model progress through correctness in constrained exercises, where the target response is predefined. In dialogue-based tutors, by contrast, learners can respond appropriately in many ways, making evidence of progress harder to interpret. This raises a learner-modeling problem: determining whether learner production provides useful evidence of progress and how they might support adaptation. We address this problem using pilot data from an LLM-based dialogue system for English grammar practice, comprising 40 pre- and post-test tasks, treatment interactions, and 2,406 learner messages. We propose a coding scheme for learner production in dialogue and explore how the resulting evidence types can support adaptive decisions. Findings show that simple use traces of the target constructions were informative for predicting independent productive use in the next task, while finer-grained evidence helped interpret the robustness of learners' developing productive control. We discuss implications for adaptive grammar-based dialogue systems that use learner production to support communicative practice.

## 1 Introduction

With the rise of large language models (LLMs), dialogue-based computer-assisted language learning (DB-CALL; [Hou and Min, 2026](#)) has become an increasingly active area of research ([Du and Reynolds, 2025](#); [Panagiotidis, 2024](#)). For grammar practice, these systems are especially promising, as they can support interactive exchanges in which learners can encounter target constructions in context, be encouraged to use them in their own responses, and receive immediate feedback.

This makes LLM-based dialogue systems well suited to support productive grammar use, understood here as a learner's ability to use the grammatical structures of a second language to express their

own meanings ([Larsen-Freeman, 2003](#)). However, making such practice adaptive remains difficult: learners differ in what they know, how they use language, and how much practice they need ([Malovrh and Lee, 2023](#); [Ruiz et al., 2023](#); [Skehan, 1991](#)).

Adaptive dialogue-based tutors, therefore, need systematic ways to infer a learner's state from complex data and use this evidence to decide what kind of support or practice should come next ([Katz et al., 2021](#)). In structured practice environments, such as exercise-based tutoring systems, a specific response is usually expected, and performance can often be evaluated through pattern matching. In dialogue, however, a learner may either respond appropriately without using the expected grammatical pattern, use it only after support, reuse it only in a familiar context, or, ideally, produce it independently across contexts. The challenge is thus not just to detect whether the expected pattern is present or not, but to interpret what it reveals about the learner's developing grammatical control.

In this paper, we investigate this issue in the context of Aisla ([Chen et al., 2022](#)), a task-based dialogue system for targeted grammar practice. The current version of the system has been designed to model and elicit a given target construction in every turn. As a next step, we aim to support adaptive practice, which requires a better understanding of how learner progress can be modeled from interaction data. We therefore ask: what aspects of learner production in dialogue provide useful evidence of productive grammar use, and do different evidence types support different modeling goals?

Drawing on pilot data comprising 2,409 learner messages, we propose and explore a coding scheme for learner production in dialogue, examining how its categories function as candidate signals of independent productive use, partial evidence of control, and continued need for practice or support.

## 2 Background

Usage-based approaches to second language acquisition (SLA) view language knowledge as organized around constructions: learned form-function mappings ranging from morphemes and specific expressions to abstract grammatical patterns (Behrens, 2009; Bybee, 2006; Schmid, 2015). From this perspective, learning involves strengthening associations between forms and functions through repeated exposure and use, making constructions more accessible and stable over time (Suzuki, 2024). With growing experience, learners may also generalize beyond specific exemplars, moving from expressions such as “I walked” toward more abstract patterns such as “I [VERB]-ed” (Goldberg, 2009; Schmid, 2018).

In addition to repetition, meaningfulness has been argued to play an important role in the development of productive grammar (Lightbown, 2007; Larsen-Freeman, 2013; DeKeyser and DeKeyser, 2007). When learners use language to express meaning across situations, they are pushed beyond repeating memorized rules, toward adapting forms to communicative needs. This makes dialogue a promising setting for grammar practice as it creates opportunities to use target constructions in context. For DB-CALL systems, however, this also creates a practical challenge: If repeated, contextualized use helps build productive control, adaptive systems need principled ways to interpret learner production as evidence for pedagogical decisions.

### 2.1 Prior Work

Existing adaptive tutoring systems often model learners from structured evidence, such as response correctness or estimated mastery (Amaral and Meurers, 2007; Portnoff et al., 2021). Recent work has begun to incorporate dialogue interaction into learner modeling, mostly outside language learning. For example, Katz et al. (2021) describe a physics tutoring system that builds a learner model during dialogue and uses it to guide adaptive support. Borchers et al. (2024) combine dialogue-act classification with skill modeling in mathematics to examine which interaction types are associated with learning rates. In language learning, Cornillie et al. (2025) predict task completion in adaptive spoken dialogue systems from process data in functional L2 speaking tasks.

Together, these studies show that dialogue interaction can inform learner modeling and adaptation.

However, they do not directly address how learner production in grammar-focused dialogue should be interpreted. In particular, prior work offers limited guidance on how to interpret different patterns of learner production in grammar-focused dialogue, such as whether learners use the practiced grammatical pattern independently or only after support, and whether non-use still reflects meaningful participation or not.

## 3 System Context

The Aisla system uses LLMs to generate targeted communicative tasks in which learners are encouraged to use target constructions while engaging in meaning-focused exchange to achieve a non-linguistic goal. During these tasks, the system provides feedback with increasing explicitness, moving from recasts to explicit grammar tips, upon persistent failure to accurately use the target form given a licensed opportunity.

The system is structured over a graph-based domain model (Ribeiro-Flucht and Chen, 2026) that links forms, functions and the authentic contexts in which they can be practiced, tasks. The graph provides an explicit representation of the instructional space, allowing the system to organize practice opportunities in a way that reflects both pedagogical structure and communicative relevance.

Learner behavior is intended to update a learner-specific overlay over this graph. The purpose of that overlay is to represent the learner’s developing control over different targets and, ultimately, to support adaptive progression through the pedagogical space. Yet such updates depend on a core unresolved issue: which aspects of learner production should count as evidence.

## 4 Methodology

We analyze 2,409 learner messages produced by 20 adult A2-level learners of English with diverse first-language backgrounds, most commonly Chinese, Portuguese, and Spanish. Over three weeks, learners completed 20 ten-turn dialogue tasks in Aisla, targeting two English constructions: the present perfect and the present continuous. The tasks were practiced in an interleaved sequence.

The study included two task types. Treatment tasks followed a short Q&A dialogue format in which the system modeled and elicited the target construction on each turn, with incremental feedback when learners did not produce the target form.

The first two and last two tasks served as pre- and post-tests. These tasks used an information-gap format: they strongly licensed the target construction through context setting and resources, but did not model it or provide turn-level feedback. This allowed us to examine whether learners produced the target construction independently in a communicative context where the form was useful but not supplied by the system. More details on the task types and task-generation procedure are reported by [Ribeiro-Flucht et al. \(2025\)](#).

Learner messages were first annotated for accurate inclusion of the target construction using POLKE ([Sagirov and Chen, 2025](#)), an English Grammar Profile annotator. We then derived additional rule-based features, including recurrent target inclusion, lexical diversity, response length, and feedback uptake, defined as target inclusion immediately after feedback. An excerpt of the annotated data can be found in [Appendix A](#). All analyses were conducted in Python 3.10. The annotated data and analysis scripts can be found in the study's Github repository <sup>1</sup>.

#### 4.1 Evidence Classification

To examine how learner production may inform future adaptive decisions, each learner message was assigned to one of five mutually exclusive evidence categories. The categories were designed as a theoretically motivated operationalization of learner evidence in dialogue. They draw on three distinctions that are central to learner modeling for productive use: whether the target construction is produced at all, whether production occurs independently or with immediate support, and whether responses without the target construction nevertheless provide evidence of meaningful task engagement.

The distinction between independent and scaffolded production is motivated by research on corrective feedback, where learner responses following feedback are treated as informative but not equivalent to unassisted production ([Lyster and Ranta, 1997](#)). The distinction between initiated and sustained independent use further separates first retrieval of a target construction from continued use once the construction has become active in the dialogue ([Ellis et al., 2015](#); [Pannell et al., 2017](#)). Finally, the distinction between engaged responses not containing the target construction and underuse reflects the ambiguity of non-use in

open-ended dialogue: a learner may respond meaningfully without using the target, or they may provide a minimal response that offers little evidence for learner modeling. We therefore propose the following categories:

- **Independently initiated:** the learner produced the target construction without scaffolding after a previous learner message that did not contain the target. This was treated as evidence that the learner could initiate use of the construction without immediate support.
- **Independently sustained:** the learner produced the target construction without scaffolding after having previously used it. This was treated as evidence that the learner could continue using the construction once it had become active in the dialogue.
- **Scaffolded:** the learner produced the target construction after receiving support from the system, such as a recast or a grammar tip. This was treated as partial evidence of productive control, since the learner used the form but did so with immediate support.
- **Engaged non-target:** the learner did not produce the target construction, but gave a response with more than two words. This was treated as evidence of task engagement without target-form use.
- **Underuse:** the learner did not produce the target construction and responded with two words or fewer. This was treated as the weakest evidence for productive control, since the response provided little opportunity to observe target-form use.

In addition, we tracked lexical diversity as the cumulative number of different lemmas used within the target construction. As discussed in [Section 2](#), from a usage-based perspective, varied lexical content in a construction may indicate more robust productive control than repeated use of a single exemplar, making lexical diversity a complementary signal of constructional development.

#### 4.2 Task-Level Prediction

Our first analysis examined whether prior learner evidence could predict independent target use in the next task. This analysis was motivated by the adaptive use case of the learner model: at each task

<sup>1</sup>[https://github.com/luisards/productive\\_grammar\\_signals](https://github.com/luisards/productive_grammar_signals)

boundary, the system needs evidence for deciding whether a learner is likely to use the target construction independently, or whether further practice and support may be needed. We therefore treated next-task independent use as a short-term prediction problem.

The unit of analysis was the learner-task pair. For each of the 20 learners, performance on tasks 2-20 was predicted from all preceding interaction data, yielding 380 observations. Because this was a small pilot dataset, the two target constructions were pooled for this analysis rather than modeled separately. This allowed us to estimate whether the proposed evidence encodings carried predictive signal across target constructions, while leaving construction-specific analyses to future work with larger samples.

The outcome was binary: whether the learner produced the target independently at least once during the task, defined as any turn classified as *Independently initiated* or *Independently sustained*. At each task boundary, features were computed from all prior turns as either cumulative or recency-weighted proportions, with exponential decay ( $\lambda = 0.85$ ). *Underuse* served as the reference category.

We compared seven nested logistic regression models: an intercept-only baseline, binary target-use traces, typed evidence profiles, and typed profiles with lexical diversity, in cumulative or recency-weighted form where applicable. This allowed us to test whether richer evidence representations improved on a simple target-use trace and whether lexical diversity added predictive value. Models were evaluated with leave-one-user-out cross-validation using AUC-ROC, accuracy, balanced accuracy, and log-loss.

All models used L2-regularized logistic regression ( $C = 1.0$ ). For recency-weighted models, decay sensitivity was also tested across  $\lambda \in \{0.75, 0.80, 0.85, 0.90, 0.95\}$ . Coefficients from typed models fit on the full sample were used to interpret the relative contribution of each evidence type, including scaffolded use, engaged non-target responses, and lexical diversity.

### 4.3 Post-test Prediction

A second analysis examined which encoding best predicted independent target use on the post-treatment probe tasks, our closest measure of less scaffold-dependent productive ability. These probes were information-gap communicative tasks with no modeling of the target form and no feed-

back.

The unit of analysis was the learner-construction pair, yielding 40 post-tests (20 learners  $\times$  2 constructions). The outcome was binary: whether the learner produced the target independently on the post-test probe task for that construction (task 19 for present perfect, task 20 for present continuous).

Predictors were derived from the treatment phase only (tasks 3-18) for the corresponding construction. These included the binary target-use rate, the rates of each evidence type, and lexical diversity. Because the sample was too small for reliable multivariate logistic regression, each feature was evaluated separately using AUC-ROC. This allowed a direct comparison of simple binary traces, typed evidence measures, and lexical diversity as candidate signals of post-treatment productive use.

### 4.4 Results

Across all learner messages, non-target responses were most common. *Underuse* accounted for 33.2% of messages and *engaged non-target* responses for 32.0%. Among target-containing messages, *independently sustained* use was most frequent (14.2%), followed by *independently initiated* use (11.5%) and *scaffolded* use (9.0%). At the task level, learners produced the target independently in 68.2% of tasks overall, although this varied widely across tasks (30%-95%) and across learners.

### 4.5 Task-Level Model Comparison

Table 1 reports leave-one-user-out cross-validation results. All evidence-based models outperformed the intercept-only baseline ( $\Delta\text{AUC} > .23$ ), but differences among the substantive models were modest. The best performance was obtained by the recency-weighted typed model with lexical diversity (M6, AUC = .751), followed closely by the cumulative and recency-weighted binary models (M1: .746; M2: .741). The main finding is that a simple binary trace of prior target use captured most of the predictive signal. Typed evidence profiles did not improve performance on their own, slightly underperforming the binary models in both cumulative and recency-weighted form. Their main contribution was therefore interpretive rather than predictive.

At the same time, coefficients from the typed models revealed a clear evidence hierarchy (Table 2): scaffolded production was the strongest positive predictor of later independent use, followed by *Independently sustained* and *Independently ini-*

Model	k	AUC	Acc.	BAcc.	Log-loss
M0: Intercept only	0	.500	.682	.500	.626
M1: Binary cumulative	1	.746	.671	.523	.548
M2: Binary recency	1	.741	.703	.608	.541
M3: Typed cumulative	4	.736	.682	.500	.569
M4: Typed recency	4	.732	.682	.500	.555
M5: Typed cumulative + diversity	5	.749	.689	.576	.547
M6: Typed rec. + diversity	5	.751	.705	.608	.538

Table 1: LOUO-CV results for all seven models.

tiated production, while engaged non-target responses were close to neutral relative to underuse. Adding lexical diversity produced small but consistent gains, and cumulative and recency-weighted encodings performed similarly overall.

Rank	Evidence type	$\beta$	OR
1	Scaffolded	+2.978	19.65
2	Independently sustained	+2.739	15.47
3	Independently initiated	+1.166	3.21
4	Engaged non-target	+0.157	1.17

Table 2: Logistic regression coefficients ( $\beta$ ) and odds ratios (OR) for typed evidence categories. Coefficients are interpreted relative to underuse, which served as the reference category.

#### 4.6 Post-test Validation

Across all probe task observations, independent target production increased from 62.5% on the pre-probe to 80.0% on the post-probe. Table 3 reports the predictive value of each treatment-phase feature for post-probe independent production in terms of area under the curve (AUC) and  $p$ -value.

Feature	AUC	$p$
Lexical diversity	.979	< .001
Binary target rate	.959	< .001
Independently initiated	.957	< .001
Scaffolded rate	.893	< .001
Independently sustained	.877	.001
Engaged non-target	.510	.934
Underuse	.049	< .001

Table 3: Predictive value of features for post-test independent target production.

Lexical diversity was the strongest predictor (AUC = .979,  $\rho = +.665$ ,  $p < .001$ ), outperforming the binary target rate (AUC = .959,  $\rho = +.637$ ,  $p < .001$ ). Unlike in the task-level analysis, where it added only modest value, it was highly informative here, suggesting that lexical variation is more closely related to broader productive control than to next-task performance.

The remaining binary and typed measures were similar in predictive power, again indicating that a simple target-use trace captures most of the useful signal. Among non-target measures, underuse was a strong negative predictor ( $p < .001$ ), whereas engaged non-target was uninformative ( $p = .934$ ). The two constructions also differed: present perfect improved from 30% to 70%, while present continuous remained near ceiling (95% to 90%). Treatment-phase evidence was therefore more informative for present perfect, where there was greater room for measurable change.

#### 4.7 Conclusion

This study examined how learner behavior in LLM-based L2 grammar practice can inform learner modeling. The findings suggest that different evidence types serve different purposes: prior target use was most useful for predicting immediate next-task performance, while finer-grained evidence helped characterize the robustness of developing productive control. Scaffolded production emerged as a strong positive signal, suggesting that system modeling and timely clarification may help make the practiced pattern available for use. Lexical diversity appeared to capture broader control over that pattern, pointing to the value of grammar practice with varied lexical fillers.

These findings suggest that less constrained communicative grammar practice may be made systematically tractable for adaptive decisions. Rather than treating learner production as too noisy to support principled adaptation, the results indicate that it can be encoded in ways that are both practically usable and theoretically meaningful. These findings are especially important for language learning, where development is gradual and linguistic knowledge is strengthened and extended through use.

## Limitations

This study should be interpreted in light of several limitations. First, the sample was small, which limits the strength and generalizability of the conclusions. Second, the analysis could not fully address cross-context tracking or cleanly separate item-specific reuse from more general productive control, beyond the lexical diversity measure. Third, the logistic regression treated observations as independent rather than modeling tasks as nested within learners, which would be preferable in a larger dataset. Finally, three learners showed ceiling-level target use across tasks, leaving little variation to explain at the individual level.

Future work should test these patterns in larger samples and with more lexically open tasks, especially to clarify whether scaffolded success reflects emerging generalization or continued dependence on practiced forms.

## References

- Luiz Amaral and Detmar Meurers. 2007. [Conceptualizing student models for ICALL](#). In *User Modeling 2007: Proceedings of the Eleventh International Conference*, Lecture Notes in Computer Science, Wien, New York, Berlin. Springer.
- Heike Behrens. 2009. [Usage-based and emergentist approaches to language acquisition](#). *Linguistics*, 47(2):383–411.
- Conrad Borchers, Kexin Yang, Jionghao Lin, Nikol Rummel, Kenneth R. Koedinger, and Vincent Aleven. 2024. [Combining dialog acts and skill modeling: What chat interactions enhance learning rates during ai-supported peer tutoring?](#) In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 117–130, Atlanta, Georgia, USA. International Educational Data Mining Society.
- Joan Bybee. 2006. [From usage to grammar: The mind's response to repetition](#). *Language*, 82(4):711–733.
- Xiaobin Chen, Elizabeth Bear, Bronson Hui, Haemant Santhi-Ponnusamy, and Detmar Meurers. 2022. Education theories and ai affordances: Design and implementation of an intelligent computer assisted language learning system. In *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium*, pages 582–585, Cham. Springer International Publishing.
- Frederik Cornillie, Julie Gijpen, Sameh Said-Metwaly, Steffen Luybaert, and Wim Van Den Noortgate. 2025. [Toward adaptive spoken dialogue systems for language learning: Predicting task completion from learning process data](#). *CALICO Journal*, 42(3):413–436.
- Robert M. DeKeyser and Robert DeKeyser. 2007. *Introduction: Situating the concept of practice*, page 1–18. Cambridge Applied Linguistics. Cambridge University Press.
- Yunfei Du and Barry Lee Reynolds. 2025. [Rooted in and beyond interaction: A systematic review of interactive affordances of chatbots for language learning amidst the rise of large language models](#). *Acta Psychologica*, 259:105307.
- Nick C. Ellis, Matthew Brook O'Donnell, and Ute Römer. 2015. *Usage-Based Language Learning*, chapter 7. John Wiley & Sons, Ltd.
- Adele E. Goldberg. 2009. [The nature of generalization in language](#). *Cognitive Linguistics*, 20(1):93–127.
- Zhuohan Hou and Shangchao Min. 2026. [Dialogue-based computer-assisted language learning systems for second language speaking development: A three-level meta-analysis](#). *ReCALL*, 38(1):40–56.
- Sandra Katz, Patricia Albacete, Irene-Angelica Chounta, and 1 others. 2021. [Linking dialogue with student modelling to create an adaptive tutoring system for conceptual physics](#). *International Journal of Artificial Intelligence in Education*, 31:397–445.
- Diane Larsen-Freeman. 2003. *Teaching Language: From Grammar to Gramming*. Thomson/Heinle, Boston.
- Diane Larsen-Freeman. 2013. [Transfer of learning transformed](#). *Language Learning*, 63(s1):107–129.
- Patsy Martin Lightbown. 2007. [Transfer appropriate processing as a model for classroom second language acquisition](#). In ZhaoHong Han, editor, *Understanding Second Language Process*, pages 27–44. Multilingual Matters, Bristol and Blue Ridge Summit.
- Roy Lyster and Leila Ranta. 1997. [Corrective feedback and learner uptake: Negotiation of form in communicative classrooms](#). *Studies in Second Language Acquisition*, 19(1):37–66.
- Paul A. Malovrh and James F. Lee. 2023. [The effects of prior knowledge on explicit rule formation among l2 spanish learners receiving structured input through guided-inductive instruction](#). *Ampersand*, 10:100116.
- Panagiotis Panagiotidis. 2024. [LLM-based chatbots in language learning](#). *European Journal of Education*, 7(1):102–123.
- Justin Pannell, Friederike Partsch, and Nicholas Fuller. 2017. [The output hypothesis: From theory to practice](#). *TESOL Working Paper Series*, 15:126–159.
- Lucy Portnoff, Erin Gustafson, Joseph Rollinson, and Klinton Bicknell. 2021. [Methods for language learning assessment at scale: Duolingo case study](#). *International Educational Data Mining Society*.

- Luisa Ribeiro-Flucht and Xiaobin Chen. 2026. [The L2 network: A CEFR-aligned knowledge graph for grammar domain modeling](#). In *Proceedings of the Workshop on Structured Linguistic Data and Evaluation*, pages 148–159, Palma de Mallorca, Spain. ELRA Language Resources Association.
- Luisa Ribeiro-Flucht, Xiaobin Chen, and Detmar Meurers. 2025. [A framework for proficiency-aligned grammar practice in LLM-based dialogue systems](#). In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 978–987, Vienna, Austria. Association for Computational Linguistics.
- Simón Ruiz, Patrick Rebuschat, and Detmar Meurers. 2023. [Supporting individualized practice through intelligent CALL](#). In Yuichi Suzuki, editor, *Practice and Automatization in Second Language Research: Perspectives from Skill Acquisition Theory and Cognitive Psychology*, pages 119–143. Routledge.
- Nelly Sagirov and Xiaobin Chen. 2025. [POLKE: A system for comprehensively annotating pedagogically-oriented grammatical structure use in language production](#). Preprint.
- Hans-Jörg Schmid. 2015. [A blueprint of the entrenchment-and- conventionalization model](#). *Yearbook of the German Cognitive Linguistics Association*, 3(1):3–26.
- Hans-Jörg Schmid. 2018. [Unifying entrenched tokens and schematized types as routinized commonalities of linguistic experience](#). *Yearbook of the German Cognitive Linguistics Association*, 6(1):167–182.
- Peter Skehan. 1991. [Individual differences in second language learning](#). *Studies in Second Language Acquisition*, 13(2):275–298.
- Yuichi Suzuki. 2024. [Skill acquisition theory: Learning-to-use and usage-for-learning SLA](#). In *Usage in Second Language Acquisition*, pages 147–168. Routledge.

## A Appendix

### (a) Present perfect – Learner 95, treatment task 9

#	Message	Evidence type
1	S: Hello! What is something you have always wanted to learn?	
2	L: to ride a bike	ENGAGED NON-TARGET
3	S: Cool! I have always wanted to ride a bike, too. Have you ever tried to learn before, or is it something you have always wished to do but never started?	
4	L: i <u>have tried</u> to learn but i failed	INDEP. INITIATED
5	S: I see. Where have you practiced riding a bike?	
6	L: i <u>haven't practiced</u> since 2015	INDEP. SUSTAINED
7	S: Oh, ok. What is something else you have always wanted to try?	
8	L: i <u>have always wanted</u> to travel abroad	INDEP. SUSTAINED
9	S: That is a wonderful wish. Which country have you always dreamed of visiting the most?	
10	L: probably canada	UNDERUSE
11	S: I have visited Canada before. I hope you get to visit there one day, too. Talk to you later!	

### (b) Present continuous – Learner 22, treatment task 12

#	Message	Evidence type
1	S: Hello. I am always watching movies in my free time. What is something that you are always doing?	
2	L: I'm always <u>watching</u> short videos.	INDEP. INITIATED
3	S: I am usually working on my computer in the evenings. What are you usually doing at that time?	
4	L: I'm usually <u>doing</u> my homework and reading books in the evenings	INDEP. SUSTAINED
5	S: That sounds productive. I am often reading in the evenings too. What kind of books are you reading these days? <sup>f</sup>	
6	L: I'm <u>reading</u> a book by a Chinese female writer. The words of the book are very delicate.	SCAFFOLDED
7	S: That sounds interesting. I am currently reading a novel too. What is the writer's name? <sup>f</sup>	
8	L: Soy tea.	UNDERUSE
9	S: Oh, you mean Su Tong? I am enjoying his writing style as well. [...] <sup>f</sup>	
10	L: I think so. See you later!	ENGAGED NON-TARGET
11	S: Yes, I am thinking his stories are very powerful. Talk to you later!	

Figure 1: Annotated dialogue excerpts illustrating the five evidence types. (a) The “Discussing long-standing wishes” task, targeting the present perfect affirmative. (b) The “Describing habits” task, targeting the present continuous affirmative form.