# Design Issues in Language Learning Based on Crowdsourcing: The Critical Role of Gameful Corrective Feedback

**Frederik Cornillie**
KU Leuven – ITEC, also at imec
KU Leuven campus Kulak Kortrijk, Etienne Sabbelaan 51, 8500 Kortrijk, Belgium
frederik.cornillie@kuleuven.be

**Abstract**

Crowdsourcing has revolutionized the software market, affecting the quality, adoption and business models of consumer software applications in many domains of human behaviour. In language learning, however, its impact is still to be seen. Through the lens of the commercial application Duolingo as well as the research prototype DialogDungeon, this paper discusses corrective feedback, a design feature of (technology-enhanced) language learning environments that can be a key driver for both learning success and platform adoption, and which will equally need to be considered in the design of language learning based on crowdsourcing. We address this topic from the literature at the intersection of second language (L2) acquisition, computer-asssisted language learning (CALL), human motivation, and gamification. We conclude with a call for collaboration between educators, L2 acquisition researchers and developers of crowdsourcing-based applications.

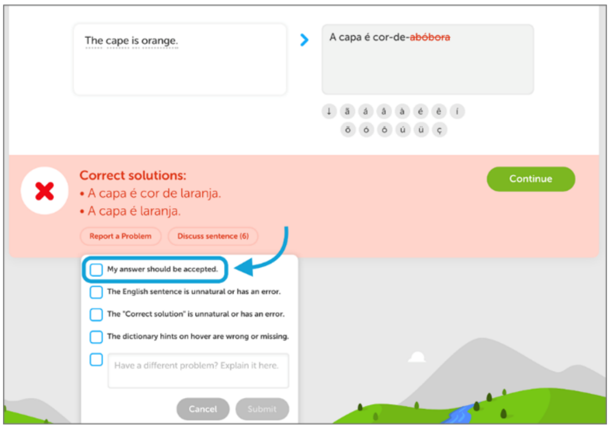
**Keywords:** digital game-based language instruction, corrective feedback, crowdsourcing

## 1. Crowdsourcing and corrective feedback in Duolingo

As a result of the Web 2.0 revolution, crowdsourcing has had a tremendous impact on the quality and adoption of many consumer software applications. Much more slowly, crowdsourcing is finding its way into research on language learning (e.g. Keuleers, Stevens, Mandera, & Brysbaert, 2015) and – arguably less effectively – into online language learning applications. The currently most popular commercial example is the gamified language learning application Duolingo, with 25 million active users on a monthly basis (Lardinois, 2018). Originally designed as a project to translate the web into every major language (von Ahn, 2013), DuoLingo is not undisputed on a pedagogical level because of its behaviourist approach to second language (L2) learning (Reinhardt, 2017; Teske, 2017; for related discussion see Cornillie & Desmet, 2016). However, its use of crowdsourcing may be useful in the L2 learning process.

On the one hand, implicit crowdsourcing of learner responses in Duolingo exercises can serve to improve the language models and learner modelling modules that among other things provide automated corrective feedback, a feature of (online) language learning environments that can be very effective when considered carefully in the instructional design process (see e.g. the meta-analysis of Li, 2010). In 2018, Duolingo organized a shared task on second language acquisition modelling, in conjunction with the 13th workshop on the innovative use of natural language processing for building educational applications (BEA) (Settles, Brust, Gustafson, Hagiwara, & Madnani, 2018). For this shared task, the company released a dataset comprising log files from millions of exercises completed by thousands of students during their first 30 days of learning on Duolingo. The goal for participants of the BEA workshop was to predict what mistakes each learner would make in the future, with a view to improving personalized instruction in the application. This shared task shows that Duolingo are actively working on leveraging state-of-the-art machine learning and psychometric techniques to improve their learner modelling and feedback generation.

From a cognitive perspective on L2 learning, this is a valuable evolution, when we consider that the effectiveness of corrective feedback depends to a great extent on individual differences (Sheen, 2011).



The cape is orange.

A capa é cor-de-abóbora

Correct solutions:
- A capa é cor de laranja.
- A capa é laranja.

□ My answer should be accepted.
□ The English sentence is unnatural or has an error.
□ The "Correct solution" is unnatural or has an error.
□ The dictionary hints on hover are wrong or missing.
□ Have a different problem? Explain it here.

Figure 1: explicit crowdsourcing in Duolingo

On the other hand, the language learning platform also involves its users in explicit crowdsourcing. For instance, learners can request that the system accepts their alternative responses, they can indicate that the language in the exercises sounds unnatural or contains mistakes, or they can discuss solutions with their peers on an online forum (see Figure 1). These activities can recruit language awareness both individually and in interaction with other L2 users, equally relevant in the L2 learning process, particularly from a (socio-)constructivist point of view (for an illustration of this approach, see Ai, 2017).

In addition to optimizing their platform through crowdsourcing, Duolingo have disclosed their interest in putting crowdsourcing to use in order to investigate L2 learning processes. Luis von Ahn, creator of Duolingo, stated that their data-driven approach and online experiments at scale can figure out "which students pick up the new concept and when", and that they can do this a lot faster than "the offline education system" (Gannes, 2014).

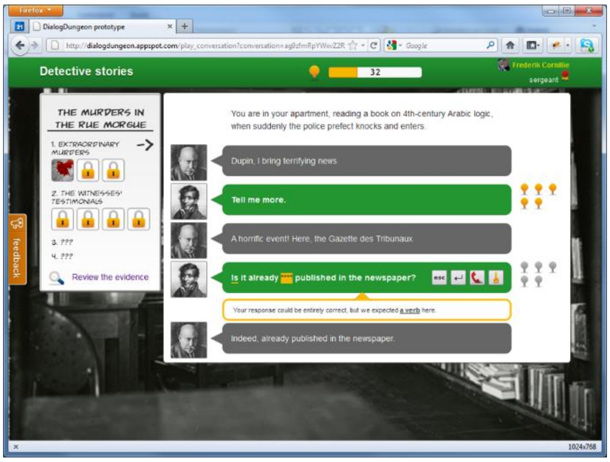
With “the offline education system”, von Ahn seems to hint at the research field of L2 acquisition. Many L2 researchers and other educational scientists will agree that this bold claim is rather simplistic – in a highly controlled environment inspired by behaviourist models of L2 learning, manipulating parameters and measuring learning outcomes is a lot easier than in more authentic language learning tasks and conditions, but the question is whether such experiments speak to ecological. Additionally, the claim seems completely ignorant of an important empirical research strand in the history of CALL, which will be discussed next.

## 2. Crowdsourcing and corrective feedback in the CALL research prototype DialogDungeon

Long before the heydays of Duolingo, CALL researchers were already exploring ideas inherent in crowdsourcing. In his keynote at the 12th International CALL Research Conference that addressed the theme “How are we doing? CALL and Monitoring the Learner”, CALL pioneer Robert Fischer reviewed studies since the early 1990s that made use of “computer-based tracking”, and argued vehemently for the analysis of tracking data with a view to “putting CALL on solid empirical footing” (Fischer, 2007). Although the scale at which these data were collected was inferior to the massive scale of data collection in contemporary applications such as Duolingo, the goals – understanding learning processes and improving CALL applications – were not fundamentally different.

More recently, Cornillie, et al.(2013) developed and evaluated a gamified dialogue-based CALL research prototype that uses crowdsourcing in language learning tasks intended to engage learners in meaningful language processing rather than in forms-focused practice (of which Duolingo is primarily an example). The goal of the project, coined *DialogDungeon*, was to design a web-based proof-of-concept application for language learning inspired by gaming, with a primary emphasis on storytelling, dialogue and learner creativity. The prototype adopted principles from the framework of Purushotma, Thorne, & Wheatley (2008) for designing video games for foreign language learning in an evidence-based way, drawing on theory and practice in L2 learning and teaching, in particular task-based language teaching (TBLT).

In the proof-of-concept, the task for the user was to solve essentially non-language-focused problems – for instance, solving a murder mystery – by using language meaningfully – for instance, asking questions as a detective. These questions and other learner responses were embedded in semi-open written activities in which the learner was required to provide a response that matched a given context. This context consisted of both the preceding and subsequent turn in the dialogue, uttered by a non-player character (see grey speech bubbles in Figure 2), as well as other specific knowledge and language related to a given dialogue or story (e.g. a bloody knife encountered in a previous scene). In addition to its task-based nature, the environment was gamified: completing dialogue turns successfully resulted in ideas, represented as light bulbs, allowing the learner to level up from constable to superintendent detective. Successful completion of dialogues yielded the learner-detective with evidence (photographs with written clues) to solve the case.



You are in your apartment, reading a book on 4th-century Arabic logic, when suddenly the police prefect knocks and enters.

Dupin, I bring terrifying news.

Tell me more.

A horrific event! Here, the Gazette des Tribunaux.

Is it already ____ published in the newspaper?

Your response could be entirely correct, but we expected a verb here.

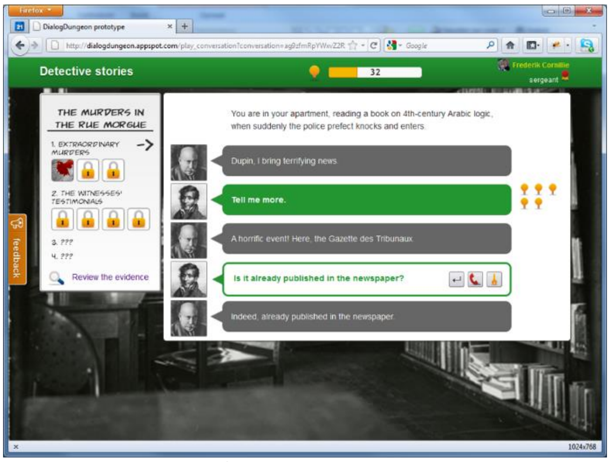
Indeed, already published in the newspaper.

Figure 2: learner completing a turn in *DialogDungeon*

The language technology that generated feedback for the learner at a given turn in the dialogue was remarkably simple, but sufficient for the task at hand, when combined with crowdsourcing. It consisted of an approximate string matching technique (based on Levinshtein edit distance, part-of-speech tagging and lemmatization) that computed the distance between the learner’s response and a set of ‘canned’ (expected) responses, which were developed by the author of the materials for each learner turn in the dialogue.

As for the ideas related to crowdsourcing, the vision of the *DialogDungeon* team was that the application had to be interesting both for language learners and native speakers. In this way, the application could collect examples of authentic language use and leverage both native speaker and learner data to enrich the dialogue models with alternative responses (both ‘correct’ responses and ‘incorrect’ ones) that were not anticipated by the dialogue author (i.e. implicit crowdsourcing from language users). In a second stage the original author of the dialogue or a teacher would annotate the collected responses for parameters like context-fit, appropriateness, and linguistic accuracy (i.e. explicit crowdsourcing from authors or teachers). A possible extension (not implemented in the prototype) was that machine learning algorithms would suggest possible scores for new responses based on their similarity to previous responses. As the application was intended to be suitable for use in instructed L2 environments, it also provided corrective feedback (based on the string matching algorithm and a set of simple rules) that consisted of highlighted (underlined) tokens and metalinguistic hints that could help learners to revise their response (see Figure 3). Finally, learners could request the responses given by their peers, ranked by frequency. This was intended as a support tool for when users got stuck in the dialogue, but the team also tinkered with the idea of using this as an entry point for having more advanced learners (or native speakers) rate their peers’ responses (explicit crowdsourcing).

An evaluation with a questionnaire showed that the majority of learners found the corrective feedback mostly useful, with a median score of 4.75 on a seven-point Likert scale, and that learners with higher prior knowledge of grammar used the feedback more often (Cornillie et al., 2013).



You are in your apartment, reading a book on 4th-century Arabic logic, when suddenly the police prefect knocks and enters.

Dupin, I bring terrifying news.

Tell me more.

A horrific event! Here, the Gazette des Tribunaux

Is it already published in the newspaper?

Indeed, already published in the newspaper.

Figure 3: corrective feedback in *DialogDungeon*

## 3. Gameful corrective feedback: potential for crowdsourcing-based CALL

One of the challenges for designers of crowdsourcing-based applications is to capture the user's attention for as long as possible, so that more (informative) user data can be collected to improve the service. Many have therefore turned to gamification, which we define as "the use of game design elements in non-game contexts" (Deterding, Dixon, Khaled, & Nacke, 2011). However, from a L2 learning and teaching perspective, it is crucial that such gamified applications are equally based on proven models of L2 learning as well as sound and widely accepted principles for L2 teaching. In other words, designers will also want user engagement with their applications to be effective, and transfer to real-life situations of communicative L2 use. Grounding the design of a crowdsourcing-based language learning application on largely discredited models of L2 learning (e.g. behaviourism) is therefore not a good starting point.

Instead, it is imperative that designers of game-based language learning applications start from the rich research literature in CALL that explores the intersections of gaming and task-based learning. Case studies in digital game-based language learning 'in the wild' (i.e. in non-instructed, informal online environments) show that such environments are particularly fecund environments for the acquisition of communicative L2 skills. In an attempt to explain this phenomenon, a number of applied linguists (e.g. Cornillie, Thorne, & Desmet, 2012; Purushotma et al., 2008) have observed that (digital) games align exceptionally well with principles of task-based language learning. First, games are all about achieving (non-linguistic) goals, such as saving the princess – pardon the masculine example. Second, in order to attain these goals, players use language (lexicogrammatical form-function-meaning mappings) meaningfully and communicatively. Language is therefore not learned intentionally, but as the by-product of engaging in tasks that are relevant to the needs of learners, which has been shown highly effective for L2 learning. Third, gaming is not play in a sandbox; it is structured play: games are structured around scenarios and mechanics. This echoes Ellis' (2003) criterial feature of a task as being a workplan. And fourth, games are intensively interactive: they react instantly to players' actions, and because players make tons of choices, this results in an endless stream of feedback. However, if designers want to translate insights from 'in the wild' case studies to formal, instructed L2 learning contexts, we need to be wary of what Larsen-Freeman (2003) called the *reflex fallacy*:

> the assumption that it is our job to re-create in our classrooms the natural conditions of acquisition present in the external environment. Instead, what we want to do as language teachers, it seems to me, is to improve upon natural acquisition, not emulate it … we want to accelerate the actual rate of acquisition beyond what the students could achieve on their own … accelerating natural learning is, after all, the purpose of formal education (p. 20)

One of the ways in which natural learning can be accelerated is by providing the learner in such task-based, meaning-focused environments with form-focused corrective feedback. Such feedback can recruit learner noticing and language awareness, focusing the learner's attention on linguistic form, which is essential for L2 development in instructed contexts. Building on empirical (including experimental) studies in the CALL literature on gaming as well as a motivational model of video game engagement grounded in Self-Determination Theory (Przybylski, Rigby, & Ryan, 2010), Cornillie (2014, 2017) elaborated a model of gameful corrective feedback that can support 'learner engagement in game-based CALL'. He defined this as learner behaviour that is driven by intrinsic motivation, that is focused primarily on language meaning and communicative use, and that involves attention to linguistic form through corrective feedback (2017). Notably, he found that gameful corrective feedback can accelerate natural L2 learning, while simultaneously stimulating intrinsic motivation, which will be associated with continued use of the environment. Designers of crowdsourcing-based CALL environments can build on this model to both enable data collection at scale and deliver effective learning experiences.

## 4. Conclusion: call for collaboration

Crowdsourcing offers exciting opportunities for L2 educators, L2 learning researchers, and developers of CALL applications. Educators will want to use crowdsourcing for at least three reasons. First, crowdsourcing allows them to personalize the learning environment for each individual learner. Second, in semi-open L2 learning tasks, it can power the generation of automated corrective feedback, necessary for accelerating natural L2 learning. Third, educators may believe in the pedagogical value of crowdsourcing because authentic language learning tasks such as storytelling are so much more interesting when the audience is actively involved, as is evident in the growing interest in fan fiction for language learning (e.g. Sauro, 2017).

Next, L2 learning researchers also have reasons to embrace crowdsourcing. It provides them with a much more fine-grained lens, combined with logistically much less demanding data collection processes, to unravel learning processes. It also allows them a methodological toolkit to study the interactions between language and its users (both 'native speakers' and 'language learners') over time, in a complex and dynamic system (De Bot, Lowie, & Verspoor, 2007).

Finally, crowdsourcing enables developers of CALL applications to launch prototypes much sooner and evaluate basic interactions at scale in order to optimize functionalities such as automated corrective feedback at a later stage. Thus, much is to be gained from an intensive collaboration between educators, researchers and developers on the topic of crowdsourcing-based CALL.

## 5. Acknowledgements

The design, development and evaluation of *DialogDungeon* was realized through the ICON project LLINGO (Language Learning in INteractive Gaming envirOnment; 2009-2011). LLINGO was funded by iMinds (now: imec) and IWT (now: Flanders Innovation & Entrepreneurship), and was carried out in collaboration with game developer Larian Studios, the Flemish radio and television broadcasting organisation VRT, Business Language and Communication Centre, and Televic Education.

## 6. Bibliographical References

Ai, H. (2017). Providing graduated corrective feedback in an intelligent computer-assisted language learning environment. *ReCALL*, *29*(May), 313–334. https://doi.org/https://doi.org/10.1017/S095834401700012X

Cornillie, F. (2014). *Adventures in red ink. Effectiveness of corrective feedback in digital game-based language learning* (Unpublished doctoral dissertation). KU Leuven (University of Leuven).

Cornillie, F. (2017). Educationally Designed Game Environments and Feedback. In S. Thorne & S. May (Eds.), *Language, Education and Technology* (pp. 361–374). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-02237-6_28

Cornillie, F., & Desmet, P. (2016). Mini-games for language learning. In F. Farr & L. Murray (Eds.), *The Routledge Handbook of Language Learning and Technology* (pp. 431–445). Abingdon: Routledge.

Cornillie, F., Lagatie, R., Vandewaetere, M., Clarebout, G., & Desmet, P. (2013). Tools that detectives use: in search of learner-related determinants for usage of optional feedback in a written murder mystery. In P. Hubbard, M. Schulze, & B. Smith (Eds.), *Learner-Computer Interaction in Language Education: A Festschrift in Honor of Robert Fischer* (pp. 22–45). San Marcos, TX: Computer Assisted Language Instruction Consortium (CALICO).

Cornillie, F., Thorne, S. L., & Desmet, P. (2012). Digital games for language learning: from hype to insight? *ReCALL*, *24*(3), 243–256.

e Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(01), 7.

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From Game Design Elements to Gamefulness : Defining “ Gamification .” In *Mindtrek 2011 Proceedings*. Tampere: ACM Press.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Fischer, R. (2007). How do we know what students are actually doing? Monitoring students’ behavior in CALL. *Computer Assisted Language Learning*, *20*(5), 409–442.

Gannes, L. (2014). Why a Computer Is Often the Best Teacher, According to Duolingo’s Luis Von Ahn. Retrieved February 3, 2019, from https://www.recode.net/2014/11/3/11632536/why-a-computer-is-often-the-best-teacher-according-to-duolingos-luis

Keuleers, E., Stevens, M., Mandera, P., & Brysbaert, M. (2015). Word knowledge in the crowd: Measuring vocabulary size and word prevalence in a massive online experiment. *Quarterly Journal of Experimental Psychology*, *68*(8), 1665–1692.

Lardinois, F. (2018). Duolingo hires its first chief marketing officer as active user numbers stagnate but revenue grows. Retrieved February 2, 2019, from https://techcrunch.com/2018/08/01/duolingo-hires-its-first-chief-marketing-officer-as-active-user-numbers-stagnate/

Larsen-Freeman, D. (2003). *Teaching language. From grammar to grammaring*. Boston: Thomson/Heinle.

Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning*, *60*(2), 309–365. Retrieved February 2, 2019, from https://doi.org/10.1111/j.1467-9922.2010.00561.x

Przybylski, A. K., Rigby, C. S., & Ryan, R. M. (2010). A motivational model of video game engagement. *Review of General Psychology*, *14*(2), 154–166. https://doi.org/10.1037/a0019440

Purushotma, R., Thorne, S. L., & Wheatley, J. (2008). *10 key principles for designing video games for foreign language learning*. Retrieved February 2, 2019, from http://knol.google.com/k/ravi-purushotma/10-key-principles-for-designing-video/27mkxqba7b13d/2#done

Reinhardt, J. (2017). Digital Gaming in L2 Teaching and Learning. In C. Chapelle & S. Sauro (Eds.), *The Handbook of Technology in Second Language Teaching and Learning* (pp. 202–216). Wiley-Blackwell.

Sauro, S. (2017). Online Fan Practices and CALL. *CALICO Journal*, *34*(2), 131–146.

Settles, B., Brust, C., Gustafson, E., Hagiwara, M., & Madnani, N. (2018). Second Language Acquisition Modeling. In *Proceedings of the NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. ACL.

Sheen, Y. (2011). *Corrective Feedback, Individual Differences and Second Language Learning*. London: Springer.

Teske, K. (2017). Learning Technology Review. Duolingo. *CALICO Journal*, *34*(3), 393–401.

von Ahn, L. (2013). Duolingo: Learn a Language for Free while Helping to Translate the Web. In *Proceedings of the 2013 international conference on Intelligent user interfaces (IUI ’13)* (pp. 1–2). New York: ACM.