Serendipitous Language Learning in Mixed Reality

Christian D. Vazquez
MIT Media Lab
75 Amherst Ave.
Cambridge, MA 02139, USA
cdvm@media.mit.edu

Megan Fu
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139, USA
meganlfu@mit.edu

Afika Nyati
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139, USA
anyati@mit.edu

Takako Aikawa
MIT GSL
77 Massachusetts Ave.
Cambridge, MA 02139, USA
taikawa@mit.edu

Alexander Luh
MIT CSAIL
32 Vassar Street
Cambridge, MA 02139, USA
aluh@mit.edu

Pattie Maes
MIT Media Lab
75 Amherst Ave.
Cambridge, MA 02139, USA
pattie@media.mit.edu

CHI'17, May 6–May 11, 2017, Denver, Colorado.

DOI string from ACM form confirmation

Abstract
Mixed Reality promises a new way to learn in the world: blending holograms with our surroundings to create contextually rich experiences that find their place in our daily routine. Existing situated learning platforms limit the learner and implicitly enforce the designer's intent by hard-coding content and predetermining what elements in the environment are actionable. In this paper, we define the framework of Serendipitous Learning in Mixed Reality as situated, incidental learning that occurs naturally in the user's environment and stems from the learner's curiosity. This framework is explored within the context of second language acquisition. We present WordSense, a Mixed Reality platform that recognizes objects in the learner's vicinity and embeds holographic content that identifies the corresponding word, provides sentence and definition cues for practice, and displays dynamic audiovisual content that shows example usage. By employing markerless tracking and dynamically linked content, WordSense enables serendipitous language learning in the wild.

Author Keywords
Augmented Reality; Mixed Reality; Language Learning; Situated Learning; Computer Vision; Serendipitous Learning

ACM Classification Keywords
H.5.1 [H.5.1 Multimedia Information Systems]: Artificial, augmented, and virtual realities

Introduction
Learning a foreign language is a dynamic process that typically begins in the classroom and ends in the real world, where we eventually validate our proficiency by interacting with the environment and others. Oftentimes, learning opportunities occur in the wild—incidentally—as we explore our surroundings, or come across new media content. More often than not, we lack the tools to harness these serendipitous learning opportunities that present themselves in the most innocuous of moments, such as during a walk through the park or the commute back home from work.

A promising way to mediate learning opportunities on the go is Augmented Reality (AR), a blooming area of research that intends to augment the user's physical world with information. Applications range from enhancing workplace performance [21] and augmenting human capability [27] to immersive recreational [20] and educational activities [26]. AR paradigms align with Constructivist learning theory [18] as a mechanism to enable situated learning outside the classroom, where the student can control the experience through exploration.

AR empowers learning experiences by exploiting the capabilities of meaningful context, which has been found to increase retention of new material [2, 5]. Moreover, situated information in AR has been shown to improve memorization and recall [6]. Despite the role of context in AR, applications that truly exploit the advantages of the learner's context are challenging due to limitations in sensing capabilities, the ability to create meaning from sensor data, and the modality through which context is made actionable.

Technical limitations in AR also hinder its potential for learning. AR on mobile phones requires context switches that increase the cognitive load on the learner and deter the capacity for immersion. Instead of looking at an augmented world, users are prompted to look at the world through the screen of a smart-phone. AR headsets often superimpose imagery on a heads up display, attempting to blend content with reality. The lack of capable depth sensing, computer vision, and optics technologies on AR devices has resulted in a plethora of artificial experiences that pragmatically offer little more than a hands-free portable display. Emerging platforms for Mixed Reality(MR), such as Microsoft Hololens [19], Meta 2 [4], and Magic Leap [14] promise what AR lacked so far: a seamless blend between the real world and virtual information.

In this work, we propose the framework of Serendipitous Learning within MR. That is, situated learning experiences that occur incidentally, and empower the learner's intent in the world. We present WordSense, a Serendipitous Learning MR platform for Second Language(L2) acquisition that automatically recognizes objects in the wearer's vicinity, and annotates them with their corresponding vocabulary words. The capabilities of the system are discussed in terms of the different embedding modalities: text, 3D models, audio, and video. The work closes with a discussion of the challenges and opportunities within the scope of the proposed framework, noting potential areas of future work and study.

Related Work
A number of works have focused attention towards AR as a platform for situated language learning. Goodwin-Jones [8] offers an overview of the emerging AR technologies in the language learning community within the last decade. We focus our discussion on two modalities for AR content: place-based and object-based.

Place-based AR leverages the importance of the user's current locality to deliver relevant and actionable materials in places that hold educational potential. The HELLO platform presented by Liu et al. [16] uses QR codes to enable tours that help students learn English as a foreign language (EFL). Similarly, Mentira [10] takes students to Los Griegos neighborhood in Albuquerque, New Mexico to solve a murder mystery while they learn Spanish. This AR experience blends location cues, cultural exposure, and collaborative activities to engage students in situated learning through gamification—a recurrent element in the field of AR Learning [15]. Although the aforementioned works can deliver powerful learning experiences they are limited to the locations for which they have been engineered; a shortcoming that we address in our serendipitous platform.

Object-based AR aims to use more granular information in the user's vicinity, often relying on cameras and sensors on the device to use objects around the learner as meaningful context. Despite existing work on object recognition and tracking in AR [7, 25], most language learning projects still focus on marker-based solutions. Wagner [29] presents a system that embeds flashcards with 3D models to create associations between Kanji characters and concepts. Hsie et al. [12] developed an augmented book with pop-up content to enhance the retention of new words. Santos et al. [24], recently introduced a platform for situated vocabulary learning that embeds text, 3D models, and audio content on objects in the learner's vicinity.

Unlike the approach we present in this paper, these aforementioned platforms rely on predefined content and marker based tracking to deliver limited learning experiences that implicitly enforce the system architect's intent by directing or influencing the learner's actions. Horneker et al. [11] found in a study that limiting interactions in AR to enforce the designer's expected usage model can lead to frustration and loss of engagement. Limiting the actionable elements in the learner's experience precludes curiosity and exploration as motivation—powerful elements that can lead to more effective learning. Greenwald et al. proposes a crowdsourcing solution that allows for dynamic and markerless language learning in TagAlong [9], which leverages a remote companion that can annotate the user's environment to enable learning and remote collaboration. However, their platform limits learning to instances where a suitable companion is available, while we provide an autonomous solution.

Serendipitous Learning in Mixed Reality

Prior work in the area of AR learning tools for language learning often relies on hard-coded content. That is, the learner is subjected to pre-established scenarios that lead to a pseudo incidental learning experience. However, outside of these carefully crafted instances, the developed technologies cannot accommodate for the learner's context that occurs naturally throughout their daily life. We call this pseudo incidental learning because it does not really occur as a fortuitous accident or unknowingly, but as a staged episode of events of which the learner is aware of. Many AR applications require the use of markers to identify places on which to embed learning content, further breaking the illusion that the learning experience happens "incidentally".

WordSense focuses on what we call Serendipitous Learning (SL) experiences in Mixed Reality. Serendipitous Learning refers to learning that occurs naturally in the learner's environment and stems from the user's curiosity; content is not fabricated or hardcoded, but instead is dynamically linked as an extension of the unique particularities of the learner's surroundings. SL is therefore situated, incidental, and aligned with pragmatics of constructivist learning the-

Figure 1: Objects are identified automatically and embedded with vocabulary words.

Figure 2: 3D models can be embedded on text to reinforce associations.

ory. We highlight three interrelated features that define SL experiences in Mixed Reality: contextual affinity, uninhibited curiosity, and dynamic content linking.

Contextual Affinity
Serendipitous Learning in MR requires the system to understand the user's environment and make it actionable as a means to enhance learning. Furthermore, the embedding is explicitly anchored in reality, placed in 3D space in such a way that it establishes a clear association between reality and content. This allows the experience to be engaging and immersive. Most importantly, it allows users to learn in real context; that is, it can establish powerful associations between objects or situations and the learning materials.

Uninhibited Curiosity
Serendipitous experiences endorse uninhibited curiosity. Instead of using markers/cues that limit and guide learning instances, the process should allow the learner to decide and explore things that truly interest him/her without feeling inhibited by the architect's intent. This allows the experience to be truly incidental as opposed to choreographed. The learner should experience the notion or illusion thereof that any object or event in the environment is a potential source of learning material as a means to diminish the learning tool's influence on the learners intent or actions. Therefore, experiences are learner-centered, where interaction is initiated by the learner's autonomous exploration of the world.

Dynamically Linked Content
We define Dynamically Linked Content (DLC) as referenced multimedia information that is linked through elements in the learner's environment. This means that learning material is not generated or manipulated by the application itself, but queried using context. DLC observes an interesting characteristic: the retrieved content is unknown—even to the architects of the learning experience—offering an element of surprise that can act as a powerful retrieval cue and fuel the learner's motivation to explore. Furthermore, the content is abundant—an extension of the web's pool of knowledge, re-purposed for learning.

WordSense: Vocabulary Learning in the Wild

In this section we describe WordSense, a Mixed Reality platform developed to facilitate dynamic, markerless embedding of content on physical objects for vocabulary learning. WordSense aims to enable Serendipitous Learning experiences for vocabulary learning by harnessing the effects of situated content on memory and association of new words. The system was prototyped using Microsoft Hololens. The client connects to a remote server hosted in the Amazon Web Services platform to obtain dynamically linked content from multiple sources.

Object to Word
Within the scope of WordSense, the primordial context consists of the objects within the learner's vicinity. An image of the target object is captured using the Hololens' front facing camera. The image is then forwarded to the Google Cloud Vision API (GCV), which offers a series of image analysis services. Using environment meshes generated by Hololens' aggregated depth sensing data, we can display the corresponding vocabulary word embedded directly on top of the identified object. Although GCV returns text in English, nouns can be translated effectively by querying services such as Google Translate to target multiple languages.

Word to Object
Optical Character Recognition (OCR) is also provided by GCV. A learner can observe written content in their environment and scan it. The obtained text is then used to query an open sourced database of 3D models [28]. These mod-

Figure 3: Sentences embedded on the target object.

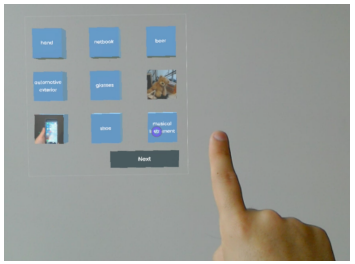

Figure 4: Review Interface to practice new vocabulary.

Figure 5: Object acts as a hub for multimedia embeddings.

els are then processed in the remote server and delivered to the Microsoft Hololens in the form of 3D meshes which are subsequently processed to generate 3D content on the fly directly in front of the text.

Sentences and Definitions
Word embeddings can be further enhanced by dynamically linking to an example sentence and definitions database. A number of definitions and sentences databases are available to the public in multiple languages. In this project, we select entries arbitrarily from a pool of sentences available in the Tatoeba database [23].

Video Clips
After an object is identified and embedded with new vocabulary, WordSense can fetch a clip that portrays the usage of the word within cinematic content. A database [17] that contains short video clips of famous movie quotes is queried using the word to identify the time where the word is spoken. The clip is fetched and streamed above the object. When multiple clips exist for a particular vocabulary word, the content is randomized.

Audio and Speech
Speech synthesis APIs for multiple languages are available. Therefore, connecting the recognized word to its pronunciation is a straightforward process. Furthermore, we explored several databases that contain recorded pronunciations of words in different languages. As a result, WordSense allows learners to hear and practice the pronunciation of newly encountered vocabulary.

Review Interface
Every object that is targeted by the WordSense application is stored for future on-the-move review. An image thumbnail is stored alongside the corresponding vocabulary word in a schema-less database solution offered by Google's Firebase platform. A learner can access this content in a flashcard fashion to reinforce retention of new vocabulary.

Discussion

The effect of glosses on second (L2) language vocabulary acquisition has been explored in many works [30, 13, 1], which generally agree that providing meaningful multimedia annotations within the context of reading material has a positive effect on the retention of new words. Dual Coding Theory [3], supports this notion by arguing that information is encoded in verbal and non-verbal cognitive processes. Encoding information in multiple mediums (e.g. images and text) and strengthening referential connections between these formats increases the probability of the learner recalling new vocabulary [22].

Mixed Reality embeddings can be thought of as an analogue of glosses within reading materials. Normally, a gloss consists of a definition associated to a new concept (vocabulary word) and links to a relevant context (sentence in reading material). In Mixed Reality, the embedded vocabulary is connected to its conceptual definition (real world object), establishing a strong association of the word within the user's reality (relevant context).

WordSense's Video Clip embeddings are a good example of serendipitous learning in Mixed Reality. Since object recognition is employed, the system has no requirement for markers to identify actionable elements in the learner's environment. Given the plurality of video content in the web that is captioned, obtaining video that contains usage of the word is straightforward and creates the illusion that any object in the environment is a source of learning material.

Many challenges exist within the framework of Serendipitous Learning in Mixed Reality. Context awareness is still limited by machine learning and sensor technologies, hin-

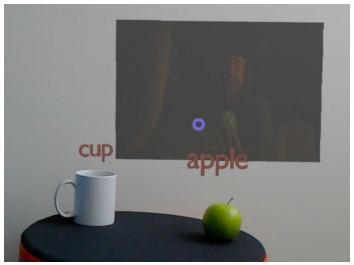

Figure 6: Video from famous movies is dynamically fetched to show usage of the encountered word.

dering the capacity for contextual affinity under certain circumstances. Within WordSense, limitations of object recognition manifest in two ways: an object is labeled incorrectly or an object is classified loosely (e.g. a cat might be identified as a mammal). A temporary solution to these problems is to display the associated words in both L1 and L2 to allow the learner to decide whether or not to trust the application. Because content is linked dynamically, the system might present embeddings that are not relevant—or worse—confuse learners by presenting erroneous content. Furthermore, since content is hosted remotely, latency can be a problem under unfavorable network conditions.

Serendipitous Learning's shortcomings are a tradeoff. The techniques presented in this work offer a more flexible approach that is scalable in terms of content and offers a truer modality for incidental, situated learning. Nevertheless, this flexibility comes at the cost of robustness. SL experiences are not meant as a substitute for traditional learning methods, but instead should be complementary to traditional curricula.

Conclusion

In this work we presented the framework of Serendipitous Learning as learning that occurs in the world, without predefined content, and uninhibited by the architect's intent. WordSense was introduced as a Mixed Reality application that enables serendipitous learning experiences for second language vocabulary learning. By combining Object Recognition with the depth sensing capabilities of Microsoft Hololens, we were able to embed real world objects with their respective words without the need for markers or predefined content. We explored how object recognition allows us to embed real words with audio, 3D models, video, and text that enables the learner to create meaningful associations between new vocabulary and concepts. Finally, we discussed the affordances and challenges of SL through the scope of WordSense, highlighting the potential to enable learner-centered activities in the wild.

Future Work

Moving forward, a series of user studies measuring the acquisition of new vocabulary using traditional and SL modalities would allow us to understand the advantages and shortcomings of the proposed framework. The capabilities of WordSense should be measured in depth to identify the rate of failure in term of three factors: inability to identify objects, retrieval of incorrect content, and retrieval of content that matches a word but not its intended meaning. A user model should be implemented to track the history of encountered vocabulary, allowing the system to provide sentences or cues that build on learned words to teach more elaborate concepts. Finally, a body of work remains in understanding how Serendipitous Learning in MR can enable more modalities for interactivity by introducing social and collaborative elements.

Acknowledgments

This work was a collaboration between the MIT Global Languages and Studies, MIT Media Laboratory, and Kanda University of International Studies (KUIS), Japan. We would like to thank KUIS for its continuing support of this project. We also thank Louisa Rosenheck, Mina Khan, and Sangwon Leigh for insightful discussions regarding the topics presented in this work.

References

[1] Samir AL JABRI. 2009. The effects of L1 and L2 glosses on reading comprehension and recalling ideas by Saudi students. (2009).

[2] Benedict Carey. 2014. *How We Learn: The Surprising Truth about When, where and why it Happens*. Pan

Macmillan.

[3] James M Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational psychology review* 3, 3 (1991), 149–210.

[4] Meta Company. 2017. Meta. (2017). https://www.metavision.com/.

[5] Matthew H Erdelyi and Jeff Kleinbard. 1978. Has Ebbinghaus decayed with time? The growth of recall (hypermnesia) over days. *Journal of Experimental Psychology: Human Learning and Memory* 4, 4 (1978), 275.

[6] Yuichiro Fujimoto, Goshiro Yamamoto, Hirokazu Kato, and Jun Miyazaki. 2012. Relation between location of information displayed by augmented reality and user's memorization. In *Proceedings of the 3rd Augmented Human International Conference*. ACM, 7.

[7] Stephan Gammeter, Alexander Gassmann, Lukas Bossard, Till Quack, and Luc Van Gool. 2010. Server-side object recognition and client-side object tracking for mobile augmented reality. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 1–8.

[8] Robert Godwin-Jones. 2016. Emerging Technologies Augmented Reality and Language Learning: From Annotated Vocabulary TO Place-Based Mobile Games. *Language Learning & Technology* 20, 3 (2016), 9–19.

[9] Scott W Greenwald, Mina Khan, Christian D Vazquez, and Pattie Maes. 2015. TagAlong: Informal Learning from a Remote Companion with Mobile Perspective Sharing. 12th International Conference on Cognition and Exploratory Learning in Digital Age 2015 (CELDA).

[10] Chris Holden and Julie Sykes. 2012. Mentira: Prototyping language-based locative gameplay. In *Mobile Media Learning*. Springer-Verlag, 111–130.

[11] Eva Hornecker and Andreas Dünser. 2009. Of pages and paddles: ChildrenâĂŹs expectations and mistaken interactions with physical–digital tools. *Interacting with Computers* 21, 1-2 (2009), 95–107.

[12] Min-Chai Hsieh and Hao-Chiang Koong Lin. 2006. Interaction design based on augmented reality technologies for English vocabulary learning. In *Proceedings of the 18th International Conference on Computers in Education*, Vol. 1. 663–666.

[13] Jan H Hulstijn, Merel Hollander, and Tine Greidanus. 1996. Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The modern language journal* 80, 3 (1996), 327–339.

[14] Magic Leap Inc. 2017. Magic Leap. (2017). https://www.magicleap.com/.

[15] Eric Klopfer and Kurt Squire. 2008. Environmental DetectivesâĂŤthe development of an augmented reality platform for environmental simulations. *Educational Technology Research and Development* 56, 2 (2008), 203–228.

[16] T-Y Liu. 2009. A context-aware ubiquitous learning environment for language listening and speaking. *Journal of Computer Assisted Learning* 25, 6 (2009), 515–527.

[17] Bill MacDonald. 2017. Quotacle. (2017). http://quotacle.com/.

[18] Jorge Martín-Gutiérrez, José Luís Saorín, Manuel Contero, Mariano Alcañiz, David C Pérez-López, and Mario Ortega. 2010. Design and validation of an augmented book for spatial abilities development in engineering students. *Computers & Graphics* 34, 1 (2010), 77–91.

[19] Microsoft. 2017. Microsoft HoloLens. (2017). https://www.microsoft.com/microsoft-hololens/en-us.

[20] Nyantic. 2017. Catch pokemon in the real world with pokemon go! (2017). http://www.pokemongo.com/.

[21] SK Ong, Y Pang, and AYC Nee. 2007. Augmented reality aided assembly design and planning. *CIRP Annals-Manufacturing Technology* 56, 1 (2007), 49–52.

[22] Allan Paivio. 2006. Dual coding theory and education. In *The Conference on Pathways to Literacy Achievement for High Poverty Children*. 1–20.

[23] Tatoeba project community. 2017. Tatoeba. (2017). https://tatoeba.org/eng.

[24] Marc Ericson C Santos, Takafumi Taketomi, Goshiro Yamamoto, Ma Mercedes T Rodrigo, Christian Sandor, Hirokazu Kato, and others. 2016. Augmented reality as multimedia: the case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning* 11, 1 (2016), 1.

[25] Rodrigo LS Silva, Paulo S Rodrigues, Diego Mazala, and Gilson Giraldi. 2004. *Applying Object Recognition and Tracking to Augmented Reality for Information Visualization*. Technical Report. Technical report, LNCC, Brazil.

[26] Kurt D Squire and Mingfong Jan. 2007. Mad City Mystery: Developing scientific argumentation skills with a place-based augmented reality game on handheld computers. *Journal of Science Education and Technology* 16, 1 (2007), 5–29.

[27] Thad Starner, Steve Mann, Bradley Rhodes, Jeffrey Levine, Jennifer Healey, Dana Kirsch, Rosalind W Picard, and Alex Pentland. 1997. Augmented reality through wearable computing. *Presence: Teleoperators and Virtual Environments* 6, 4 (1997), 386–398.

[28] Exocortex Technologies. 2017. Clara.io: Online 3D Modeling, 3D Rendering, Free 3D Models. (2017). https://clara.io/.

[29] Daniel Wagner and Istvan Barakonyi. 2003. Augmented Reality Kanji Learning. In *Proceedings of the 2Nd IEEE/ACM International Symposium on Mixed and Augmented Reality (ISMAR '03)*. IEEE Computer Society, Washington, DC, USA, 335–. http://dl.acm.org/citation.cfm?id=946248.946816

[30] Makoto Yoshii. 2006. L1 and L2 glosses: Their effects on incidental vocabulary learning. *Language Learning & Technology* 10, 3 (2006), 85–101.