Contextualizing Language Learning in the Digital Wild: Tools and a Framework

David Wible
Dept. of English
Tamkang University, Taiwan,
dwible@mail.tku.edu.tw

Chin-Hwa Kuo
CAN Lab., Dept. of CSIE,
Tamkang University, Taiwan,
chkuo@mail.tku.edu.tw

Nai-Lung Tsao
CAN Lab., Dept. of CSIE,
Tamkang University, Taiwan,
beaktsao@mail2000.com.tw

Abstract

A premise of this paper is that there are distinctive qualities of the domain of language which render learning a language uniquely suitable for a radical contextualization within digital environments. The paper thus describes and illustrates an approach to supporting foreign language learning ubiquitously in unrestricted networked environments. The two tools presented focus on vocabulary learning. The Collocator tool detects and highlights collocations (such as 'prescribe medicine' or 'stiff competition') in real time on any web page the user is viewing. The user can select any of the highlighted collocations for focused attention, activating a 'push' mechanism that will provide repeated examples of the collocation over the ensuing days. Word Spider allows users to select unknown words in any web text and it responds by finding semantically related words in its context, automatically annotating these, exploiting them as contextual clues to the meaning of the targeted unknown word.

1. Introduction

The success of any approach to digital learning will rest to a great extent upon finding a match between the nature of the particular domain of learning on the one hand (say, math, physics, history, or language) and the details of the digital environment provided for that learning on the other. A premise of this paper is that there are distinctive qualities of the domain of language which render learning a language uniquely suitable for a sort of radical contextualization or radical embedding within digital environments. The goal of the paper is to describe and illustrate such a radical embedding and two novel tools which implement it. The point of departure for this approach is a language learning platform developed and implemented over the past four years called IWiLL (Intelligent Web-based Interactive Language Learning)[1]. What we present here is a novel extension of IWiLL from its original design as an autonomous web-based platform to a new diffused platform-independent architecture that provides personalized English learning support ubiquitously wherever users browse on the web. Accordingly, this novel implementation is titled UWiLL (Ubiquitous Web-based Interactive Language Learning). Specifically, rather than focusing upon how to construct an autonomous language learning platform [1][2], or how to design digital content for language learning, we propose the alternative approach of embedding language learning within existing noisy online environments. Taking the case of English learning, we show that the central challenge of such an approach is how to take the existing online English environments that users freely browse (for example, news, sports or entertainment websites) and transform them in real time into environments that enhance these users' English learning. In what follows, we describe and motivate this radical contextualization of language learning, elucidate the properties of language which make it particularly suitable for this sort of radical

[1] IWiLL has been used by 196 schools, by 643 different teachers, 23,444 students and 2,075 independent learners. Teachers have authored 2,470 web-based lessons with the system's authoring tool. A learner corpus automatically constructed from the use of IWiLL (English TLC) has archived over 29,000 English essays consisting of a total of almost three million words of machine-readable running text written by Taiwan's learners using the IWiLL writing platform. These essays have been marked digitally by teachers on that platform with over 47,000 comments, each comment indexed to the marked error in the student essays. Investigations into this corpus along with the archive of teachers' comments have led to novel research methodologies and insights into Taiwan's learners' English. See [1] and [2] for a detailed description of IWiLL.

Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT'04)
0-7695-2181-9/04 $20.00 © 2004 **IEEE**

embedding, and present two novel tools that implement this approach of embedding language learning in the digital wild.

2. An Approach to Contextualizing Language Learning in Noisy Digital Environments

One of the premises of this paper is that, in the domain of learning a second or foreign language, certain distinctive properties of the domain of language and the nature of language learning (sketched below) converge to yield unique possibilities for seamlessly embedding language learning within a learner's overall experiences in digital environments.

2.1. Alleviating the Content Bottleneck

Anyone involved in digital learning design in recent years is aware that one of the central obstacles to fulfilling the widely touted potential of the field is a content bottleneck. Supporting embedded language learning ubiquitously on the web provides a way of alleviating the digital content bottleneck since it consists of accompanying users in their unrestricted online activity with tools to enhance these noisy environments in real time in ways that support English learning. This approach stands in contrast to the more common traditional practice of designing digital content or even autonomous digital environments specifically for language learning.

2.2. If we build it, will they come?

Even granting the highly optimistic assumption that sufficient digital content can be created to relieve the content bottleneck, this is no guarantee that the targeted users will make use of this content. A common complaint of platform and content designers who have already built formidable sites with high quality content is the disappointingly low level of usage these materials receive. The alternative approach advocated here is that rather than assuming "if we build it, they will come," we assume that the intended users may not come and that it is worth attempting instead to follow these users wherever it is they happen to be going on the web. In the case of language learning, this fosters language learning within authentic contexts freely selected by learners wherever they go on the web rather than within contrived contexts imposed upon them. While it would be difficult to make a case that other domains (for example, physics, math, or history) could be learned within web environments that unrestricted learners freely browse, the domain of language, we suggest, lends itself uniquely to such contextualization. It is worth briefly describing, then, what it is about language that affords this alternative possibility.

2.3. Distinctive Properties of Language as a Learning Domain

The two motivations described above (alleviating the content bottleneck and creating authentic and learner-centered contexts for language learning) are afforded because of certain unique characteristics of language as a learning domain. Three of these properties are described briefly here.

First, language is ubiquitous. Unlike the case with, say, physics or math, in order to be in an environment suitable for learning a language, one need not enter a 'language classroom' nor a website designed for 'language learning'. In fact, the driving assumption of the proposed ubiquitous support for networked language learning is this: Every 'English' environment is a potential 'English learning' environment.

A second fact which sets language apart from other domains of learning is that the cognitive mechanisms of language learning differ from those of other learning domains. While it is impossible to master physics or math or geography or history without acquiring conscious knowledge of the content of these domains, it is indeed possible to master a language without conscious knowledge of the so-called rules of that language, or at least certainly without conventional instruction. This is true for virtually everyone when it comes to his or her native language. It is also true for many second language learners who have acquired their second language outside a classroom setting..

A third property that makes language unique among learning domains is that its central purpose is to communicate, to convey meaning. Thus, embedding the learning of a target language within online contexts where users are attempting to use the language for actual communication (for example, to get information) is as well-motivated as embedding the learning of swimming within a swimming pool as opposed to a 'dry' decontextualized' classroom.

3. The Tools: Collocator and Word Spider

Word Spider and Collocator are tools aimed at fulfilling our goal of ubiquitous language learning support. These two tools support vocabulary acquisition in the context of free unrestricted web

browsing. Both tools are accessed via a toolbar that appears on the user interface during all web browsing. Each of these two tools addresses a different type of challenge in vocabulary learning in context. The language support provided by each tool is described and illustrated in detail in what follows.

3.1 Collocator: A Tool for Detecting Collocations in Context

Collocations constitute one of the most persistent areas of difficulty for learners in acquiring second language vocabulary. Our research team has been looking at learner collocation errors and at computational aspects of collocation detection since 2000[2][3][4][7]. Liu[7], for example, analyzed a range of miscollocations from a learner corpus (English TLC) and uncovered a dramatic concentration of verb-noun miscollocations (e.g., **pay…time* vs *spend…time*) compared to other part-of-speech combinations. Moreover, she found that in over 97% of these VN miscollocations, it was the verb rather than the noun which was incorrect. These findings proved invaluable in subsequent attempts to automate the correction of miscollocations. Liu is also the first to propose a semantic approach to automating the correction of miscollocations, specifically seeking candidate replacement verbs from among verbs semantically related to the incorrect verb in VN miscollocations, using WordNet as the source of these candidates. Her hand-constructed rules covered 36 different VN combinations and achieved a precision rate of over 95% in correcting these targeted miscollocations [2]. Pilots of our broader attempts to automate miscollocation correction in general, achieved up to 85% precision rates, not sufficiently high in our estimation to embed in applications for users. In contrast, the Collocator tool described below, rather than addressing the miscollocation output produced by learners, focuses on detecting collocations in standard English input learners encounter on line, and it is ready for deployment for learners in a browser-based toolbar. Collocator's design is motivated and described below.

An apparent source of difficulty for learners in acquiring collocations is that they are idiosyncratic. For example, there would appear to be nothing in the meaning of the words involved which would predict that *make a conclusion* odd whereas *draw a conclusion* is acceptable, or that we can intensify the noun *respect* with the adjective *great* (They have *great respect* for her) but not with the near synonymous adjective *big* (*They have *big respect* for her). These restrictions illustrate the phenomenon of collocation: many words are unpredictably picky about the other words with which they can co-occur. The central motivation for our Collocator tool is that this pickiness (or 'collocability'), which learners must master, is not detectable from their direct encounters with target language input. There is nothing, for example, in the appearance of *take medicine* and *buy medicine* in the same text which would signal that the one is a collocation and the other is just a free combination. That is, nothing from these instances would indicate that the verb *take* in the collocation *take medicine* cannot be freely replaced with synonyms or other plausible verbs, such as *eat medicine*, whereas the verb in the free combination *buy medicine* can indeed be replaced by a synonym, as in *purchase medicine*. The point here is that there is nothing directly in the texts that users encounter that would indicate which phrases are collocations and must be mastered and which are just free combinations. In fact, this is precisely why computational methods for collocation detection require sophisticated statistical measures run over very large corpora (See [5] [6], inter alia.) and why learners require vast amounts of accumulated experience with the target language to acquire collocations.

The purpose of Collocator is to offer the learners collocational knowledge from a single reading experience which would otherwise have to come from massive amounts of contextualized exposure to the words involved. The approach of Collocator is to enhance reading texts that are freely selected by learners on line, highlighting for them in real time precisely those word combinations in the text that are collocations. The tool detects such collocations in the learner's text in real time by exploiting statistical word association measures on a 30-million-word portion of the British National Corpus (BNC)[10]. Combinations of words that achieve a sufficiently high association score (calculated on BNC) to constitute collocations and which co-occur within a specified window of proximity to each other in the targeted text are highlighted there as potential collocations.

Figure 1 shows the collocation *prescribe medicine* highlighted by Collocator and a pop-up text indicating its status as a collocation. A link from each detected collocation is added in real time which lists additional examples of this same collocation in order to provide users with richer and more intensive exposure to the same collocation.

An additional feature of Collocator is a personalization module which allows users to mark specific collocations that have been detected by Collocator to have these recorded in this user's personal profile. Collocator can then either

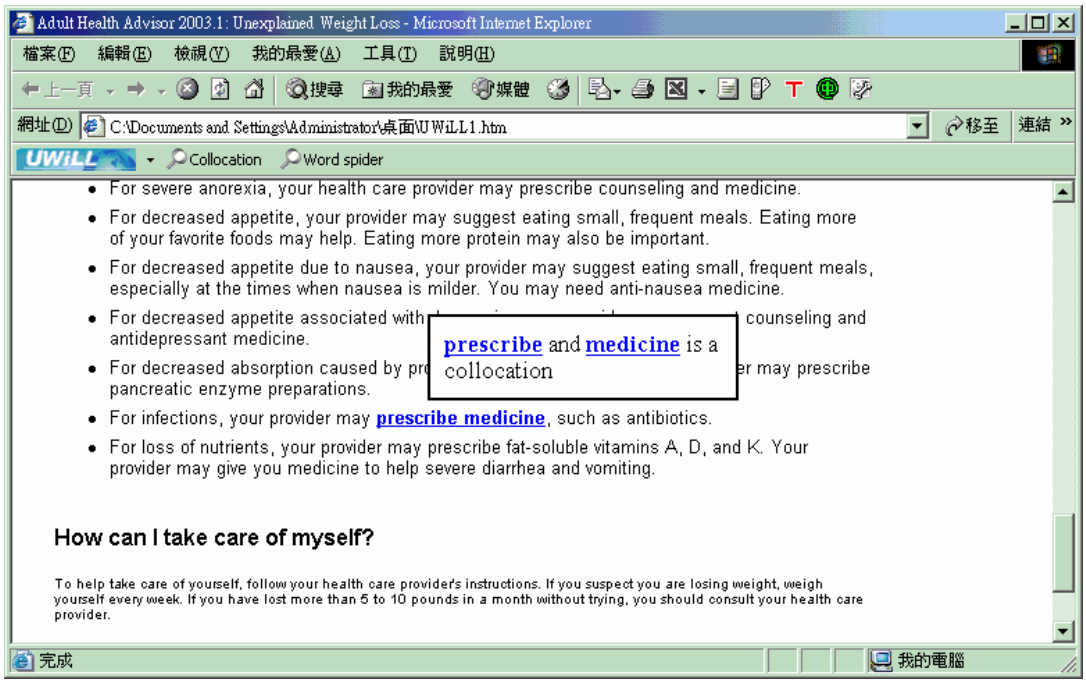

Figure 1. The collocation 'prescribe...medicine' detected by Collocator

automatically detect these high priority collocations in the web pages that the user accesses in the future to provide repeated exposure to this same collocation over an extended period of time or, rather than waiting until the user happens onto the targeted collocation in future browsing, can push contextualized examples of these high priority collocations extracted from BNC.

3.2. Word Spider: A Tool for Providing Clues to Unknown Words in Context

The second tool, Word Spider, addresses a challenge for vocabulary acquisition somewhat different from that posed by collocations. This challenge is how to deal with individual words in a text which are completely unknown to the user [8]. The function of Word Spider is to take any such unknown noun or verb encountered by a user in a web page text and, once selected by the user for Word Spider's assistance, to search the context (same web page) for any other words that are semantically related to this unknown word and which could serve as clues to the meaning of the unknown word. Word Spider then highlights these semantically related words and, with a mouse-over, provides a pop-up annotation describing the relation between the two words. Figure 2 illustrates this function, taking the word *antibiotics* as an example unknown word selected by a user for Word Spider's assistance. As the figure 2 shows, Word Spider detected the word *medicine* preceding *antibiotics* in the same sentence and highlighted it in a different color. The pop-up shows the automatic annotation given by Word Spider as a clue to the meaning of *antibiotics*: "Antibiotics is a kind of medicine." This automatic detection of surrounding clues and the automatic annotation of their relation exploits an existing lexical database, WordNet, which encodes the lexical semantic relations described above, relations such as hypernym and hyponym holding among words or, more precisely, among sets of synonyms that represent word senses [9]. The database encodes a set of lexical semantic hierarchies which we exploit to support the sort of inferencing involved in providing contextual clues to the meanings of unknown words.

It is worth noting here that while our team has the computational and lexical resources to provide simply pop-up definitions or glosses in the users' first language (Chinese) for all words in an English text encountered online, we eschew this approach for pedagogical reasons. Specifically, we are interested in not simply transmitting lexical information to learners on individual words they choose. Rather, we are interested in cultivating in learners a healthy reading strategy of seeking contextual clues to guess the meaning of unknown words. We have strong reservations concerning the effects of providing direct pop-up definitions or translations on the reading

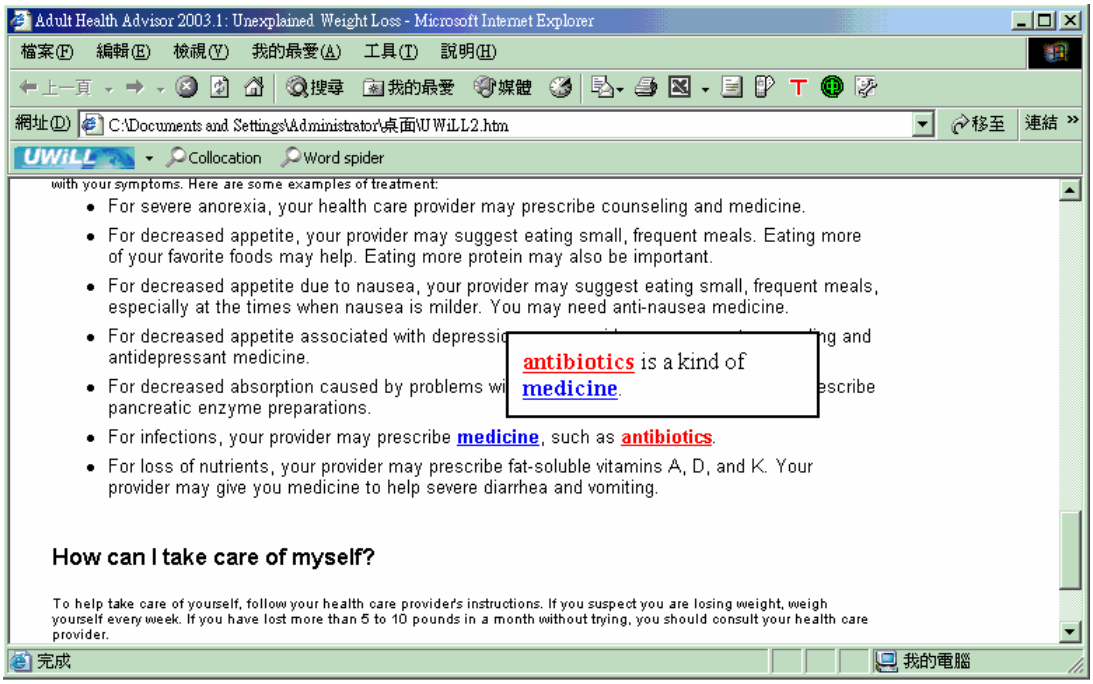

Figure 2: Word Spider detects a contextual clue to the meaning of 'antibiotics'

strategies or vocabulary learning strategies of second language learners.

If users find that the surrounding words highlighted as clues by Word Spider are too few to be helpful, or if no such clues are detected in the context by Word Spider, the user has the alternative of simply having clues provided in the pop-up. For example, even if the word *medicine* were not found in the context of the word *antibiotics*, Word Spider could still provide the annotation "Antibiotics is a kind of medicine." Polysemy (words with more than one meaning) presents a challenge for Word Spider. For targeted words that have more than one meaning, currently we simply provide both (or all) possibilities and let the user attempt to determine which is the relevant one in the case at hand. For example, if the word party were targeted by a user, Word Spider would provide the following sort of clue if none were found in context: "Party can mean a kind of event or a kind of organization. Which do you think is right here?"

4. References

[1] Wible, David, Chin-Hwa Kuo, Feng-yi Chien, Anne Liu, and Nai-Lung Tsao, "A Web-based EFL Writing Environment: Exploiting Information for Learners, Teachers, and Researchers", *Computers and Education*, 2001, vol. 37, pp. 297-315.

[2] Wible, David, Chin-Hwa Kuo, Nai-Lung Tsao, Anne Liu, and Hsiu-ling Lin, "Bootstrapping in a Language Learning Environment", *Journal of Computer-Assisted Learning*, 2003, vol 19 #1, pp. 90-102.

[3] Wible, David, Chin-Hwa Kuo, and Nai-Lung Tsao, "Improving the Extraction of Collocations with High Frequency Words", *International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugual, May, 2004 (accepted).

[4] Wible, David and Anne Liu, "A Syntax-Lexical Semantics Interface Analysis of Collocation Errors", *PacSLRF (Pacific Second Language Research Forum)*, University of Hawaii, Manoa, Hawaii, 2001.

[5] Biber, Douglas, "Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition", *Computational Linguistics*, 1993, Vol. 19 #3, pp. 531-538.

[6] Dunning, Ted, "Accucate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, 1993, vol. 19 #1, pp. 61-74.

[7] Liu, A. "*A Corpus-based Lexical Semantic Investigation of Verb-Noun Miscollocations in Taiwan Learners' English*" ,MA thesis, Tamkang University, Taipei, 2002.

[8] Prince, P., "Second language vocabulary learning: The role of context versus translation as a function of proficiency", *The Modern Language Journal*, 1996, 80(4), 478-49.

[9] Miller, G., "WordNet: An Online Lexical Database", *International Journal of Lexicography*, 1990, Vol 3, #4.

[10] http://www.natcorp.ox.ac.uk/.