行政院國家科學委員會專題研究計畫　期中進度報告

語言習得，分散式認知與學習科技的交集(2/3)

計畫類別：整合型計畫
計畫編號：NSC94-2524-S-032-005-
執行期間：94 年 05 月 01 日至 95 年 07 月 31 日
執行單位：淡江大學英文學系

計畫主持人：衛友賢

計畫參與人員：陳惠如，黃平宇

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 5 月 29 日

0 Preface

The overarching purpose of the present E-Learning Technology project is to develop and implement a novel approach to web-supported language learning which frees learners from the limitations of traditional online packaged content. We have developed an alternative approach called *UWiLL* (for ubiquitous web-based interactice language learning) which embeds our language learning tools within the learners' web browser in the form of a toolbar. The novel research challenge that our approach poses is how to make our tools sufficiently flexible and robust to provide relevant vocabulary assistance in any Web environment the learner chooses to browse. We have already developed a suite of innovative stable tools for this purpose. The paper that follows describes our latest addition to this suite of these tools, which incorporates a novel learner model and thereby enables the browser-based tools to pinpoint in real time during the users' unrestricted Web browsing specific vocabulary uses that pose particular difficulties for our learner population and to supplement these cases with specific advice concerning the users' likely troubles with this vocabulary usage. This paper will appear in *Lecture Notes in Computer Science* (SCI-extended) as a full paper from ITS 2006. The acceptance rate for ITS 2006 was 38%, and although the conference will be held in Taiwan this year, only three of the 60-some full papers are from Taiwan. Our paper is one of those three.

Abstract. One of the most persistently difficult aspects of vocabulary for foreign language learners is collocation. This paper describes a browser-based agent that assists learners in acquiring collocations in context during their unrestricted Web browsing. The agent overcomes the limitations imposed by learner models in traditional ITS. Its capacity to function in noisy unscripted contexts derives from a well-understood theory of lexical knowledge that attributes a word's identity to its contextual features. Collocations constitute a central feature type, and we extract these features computationally from a 20-million-word portion of BNC. These we are able to detect and highlight in real time for learners in the noisy Web environments they freely browse. Our learner model, derived by semi-automatic techniques from our 3-million word corpus of learner English, maps detected

collocations onto corresponding collocation errors produced by this learner population, alerting learners to the non-substitutability of words within the target collocations. A notebook offers a push function for individualized repeated exposure to examples of these collocations in context.

1 Purpose and Motivation

One of the most serious limitations in ITS is the fragility of learner models. A common consequence of a lack of robust learner models for a particular learning domain is that intelligent systems typically impose tight restrictions upon the learners. Only with such restrictions can the learner's behavior become predictable enough to enable the system to respond intelligently (and appropriately) within the scope the system's limited expertise. This is not only the case where learner models are meager or poorly articulated. Often equally limiting are highly articulated learner models because these correspondingly require highly articulated scripts to guarantee that this model can derive the needed inferences from the learner's behavior. Thus, quite generally, learner freedom and flexibility are often sacrificed as a prerequisite for expressing the system's intelligence.

Such tight restrictions are especially regrettable in foreign language learning, where the goal is to gain competence in using language to express personal meanings and to understand the meanings expressed by others in a range of contexts. Moreover, one of the richest contexts where learners can be exposed to the target language used for such authentic communication is the Web. Such rich exposure to language input offered by the Web also addresses one of the persistent limitations of traditional classroom foreign language learning: underexposure to the target language in authentic contexts. Unfortunately for system designers, the Web is correspondingly noisy and the sorts of language and contexts that the users may encounter are unpredictable. In earlier work we have referred to this environment as the "digital wild" [13]. The purpose of our recent research has been to develop an approach to designing digital tools sufficiently robust to provide context-sensitive personalized help to language learners in such environments. Here we describe and illustrate here a ubiquitous agent that provides this sort of help for vocabulary learning.

2 Approach

We refer to the overall research framework and infrastructure that we have developed under this browser-based approach with the acronym UWiLL (ubiquitous web-based interactive language learning). The tools reported in this paper build upon the infrastructure of context-aware browser-based language tools we recently developed under UWiLL [13].

In this paper we address the limitation that imperfect learner models typically translate into restrictions on the learners. We describe and illustrate an approach which retains both the learners' freedom and the system's responsiveness to them. Here our approach relies upon two fundamental ingredients: (1) a highly articulated yet computationally tractable theory of the target knowledge domain and (2) a correspondingly tractable theory of the knowledge acquisition: that is, a theory of what it takes to acquire this target knowledge. Within our way of framing the problem, once these two key ingredients are in place, the burden on the learner model eases dramatically, to the point where personalization can be achieved in these same noisy conditions with the addition of a relatively simple, straightforward learner model. In what follows we show how this is so.

3 A Computationally Tractable Model of the Domain Knowledge

A universal assumption in second language acquisition (SLA) research is that exposure to the target language is the *sine qua non* of language acquisition. Yet learners must eventually glean from this exposure a mastery of the target language system (whether consciously or unconsciously is a contested question that need not concern us here). Thus, exposure to target language is useful to the extent that the learner can distill from this experience the features of the language that must be mastered, for example, to distill from exposure to English the fact that English requires verbs to agree with their subjects in finite clauses [5][8]. We take this to be the key desideratum for our ubiquitous agent. Specifically, to function in the noise of the unrestricted Web, such an agent must be able to detect in real time within this noise whatever salient linguistic features it is designed to help the learner acquire. Our agent is viable for two reasons. We have an explicit theory of these linguistic features and we have computational tools that can extract them from noisy texts in real time.

4 Collocation and a Theory of Contextual Features

The specific domain of language learning that we target here is vocabulary learning. Thus the purpose of our agent is to help learners increase their mastery of the target language vocabulary in noisy unscripted contexts. Accordingly, our approach, sketched above, requires that our agent be able to detect within these unrestricted contexts precisely those linguistic features that govern the mastery of the target vocabulary. To achieve this, we need a machine-tractable theory of these features.

For this, we subscribe to a contextual view of words. This view assumes that words are, by their very nature, contextual creatures and that mastery of a word consists essentially in mastering the contextual features that govern that word's felicitous use. One of the most widely exploited types of contextual features of words in the computational linguistics literature is collocation (for example,[3][15]; inter alia). Thus, the salient contextual features of the target word that we exploit are the collocating words (or collocates) of that word. This underlying assumption is captured in the famous quote of British linguist J.R. Firth: "You shall know a word by the company it keeps" [4]. Essentially, the collocates of a word are "the company it keeps," that is, a word's collocates are the other words which it conventionally co-occurs with.

Here we motivate the notion of collocation as a fundamental dimension of the contextual features that make up a word's identity. Along the way we show a word's collocates to be (1) features that the learner must eventually master as a key aspect of vocabulary learning and (2) features that we can extract computationally in real time and detect under noisy condition for the learner's attention.

Following Manning and Schütze [7], we refer to the target word of interest as the *focal word* and to the words that this target word selects for its contextualized use as the *collocates* of that focal word. Hence a collocation comprises a pair of words: a focal word and one of its collocates. Part of mastering a noun, for example, is to master its collocates. Lack of this mastery leads learners to produce expressions such as *big rain, big wind, big respect* ("I have big respect for that coach"). These errors arise from inadequate collocation knowledge. Each of these three nouns, taken as different focal words, imposes different requirements on the selection of its collocates; they each require a different adjective to express the intended meaning: *heavy rain, strong wind, great respect*. Collocational knowledge is heavily idiomatic. That is, it does not readily generalize (e.g., *heavy rain* but not *heavy wind*; *strong wind* but not *strong rain*). Again, on our view, collocates as contextual features are not secondary aspects of word knowledge; they constitute the very heart of a focal word's identity. We have mentioned a central motivation of

our work to be learners' need for adequate exposure to the target language as a means of mastering the features of the target language. Here we can frame this motivation for the particular issue of learning collocations. Specifically, second language learners require sufficient exposure to vocabulary words in context in order to detect and internalize the collocates of these words.

Wang [9] provides empirical evidence that in the particular case of collocation learning in a foreign language, exposure to examples of the target collocation is an effective strategy in helping learners acquire collocations. In fact, exposure to positive examples of a target collocation was dramatically more effective than teacher corrections and comments in helping learners acquire a collocation they had misused. Thus, the central pedagogical strategy of our ubiquitous agent is to draw the learner's attention to collocations detected in their web browsing and then to supplement this highlighting with numerous example sentences containing the detected collocation.

5 The Design of the Ubiquitous Agent: Collocator

We refer to our ubiquitous agent as Collocator. The design of Collocator exploits the free Web browsing of learners to provide the intensive exposure to collocations that is required if those collocations are to be acquired. To do this, the agent detects collocations that occur in the Web pages that the learner freely browses and then, at the request of the learner, highlights any of these detected collocations in their context on the Web page. To intensify this single exposure to the detected collocation, the agent then provides, again at the learner's request, numerous example sentences containing that same collocation. A notebook function then enables the learner to select any of these collocations for future review and to store any of the example sentences provided by Collocator. Specifically, Collocator provides a "push" request which allows the learner to request repeated exposures to any of the detected collocations with example sentences over subsequent days, thus reinforcing the single exposure highlighted by Collocator during browsing.

There are two versions of Collocator that operate simultaneously: *G-Collocator* (G for Greedy or General Collocator) and *P-Collocator* (P for Picky or Personalized Collocator). G-Collocator runs on an algorithm (to be described below) that detects any word pairs that exhibit a sufficiently strong word association score and treats these pairs as collocations. P-Collocator is more selective, containing a learner model that indicates which collocations have been misused by this learner population and thus require special attention. The design

architecture of both G- and P- Collocator are described in what follows.

5.1 Components of the Browser-based Agent

The schema in Figure 1 represents the components of Collocator, the browser-based agent.

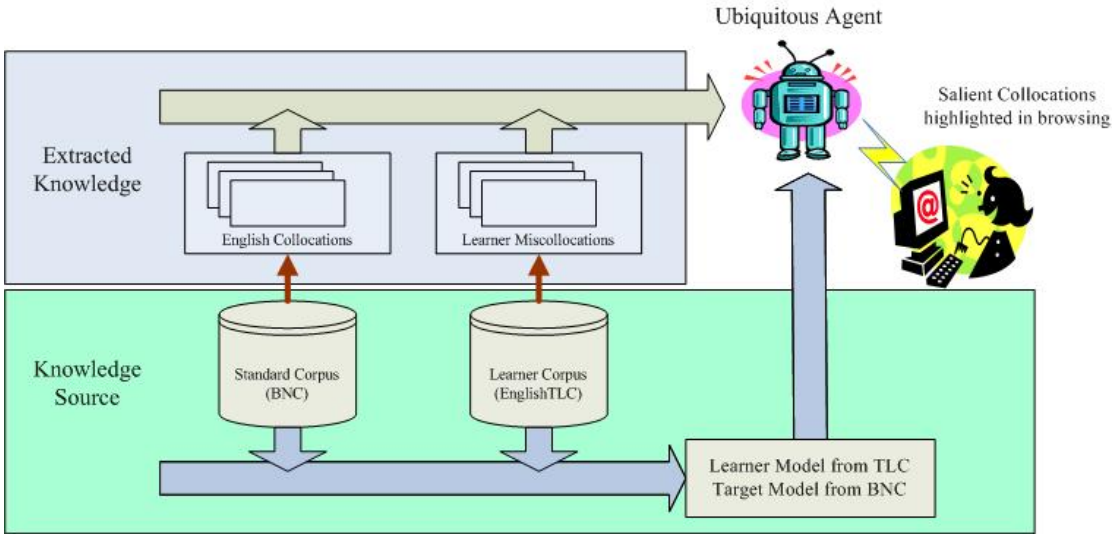

Figure 1

Here we describe this architecture schematically and then in the following section provide details of each component and its relation to the overall system.

The schema shows two levels of knowledge. The lower level contains two sources of knowledge that feed the agent; the upper level contains two counterpart sorts of knowledge extracted from these sources. The two knowledge sources represented on the lower level are (1) a standard English corpus (a 20-million-word portion of the British National Corpus, which we have re-indexed for efficient real-time collocation extraction) and (2) our corpus of learner English—*EnglishTLC* (3 million words of English running text produced by Taiwan learners).

From this lower level of knowledge sources we derive the upper level—extracted knowledge of two sorts. The first sort of extracted knowledge is our domain knowledge model consisting of English collocations. These are extracted from BNC through statistical word association measures. We use a traditional mutual information (MI) measure combined with our own variation of this which detects collocations that traditional MI underextracts (See [14]). The extracted collocations then serve as the target knowledge model—standard collocations. These are the collocations detected by G-Collocator. It has no particular learner model, but provides exposure to any collocations detected in the Web pages the user browses. Hence, G in G-Collocator suggests Greedy or General collocation detection.

The second archive of extracted knowledge represented on the upper level of the schema is the relevant learner model used for P-Collocator, for Personalized or Picky collocation detection. This model, derived from our 3-million-word learner corpus, consists of attested miscollocations produced by our population of learners. Miscollocations are errors such as *pay time* (rather than the correct *spend time*) or *learn knowledge* (rather than the acceptable *gain knowledge* or *acquire knowledge*). We use two techniques for mapping these miscollocations attested in our learner model onto the corresponding acceptable collocations found in the domain knowledge model that can be used in their place (for example, mapping the miscollocation *eat medicine* onto the correct collocation *take medicin*e). This mapping is the core knowledge deployed by P-Collocator. The specific target or correct collocations identified by this mapping are what we refer to in the schema as 'salient collocations'. In what follows, we describe the functionality of this agent as it accompanies learners in the context of their unrestricted Web browsing.

5.2 G-Collocator (GC)

As mentioned above, GC detects every valuable collocation in browsed web pages. In this respect, it is greedy or general. The collocation-extracting scheme is part-of-speech sensitive, which means we have to know the part-of-speech information of each word in browsed web pages. We train a Markov Model-based POS tagger [1] and use British National Corpus (BNC)[1] as our training data. The internal evaluation shows this tagger has 93% precision including identifying unknown words. After part-of-speech tagging, the agent uses the following equation from [14] to measure the word association score:

$$normMI(x, y) = \log_2 \frac{P(x, y)}{\left(P(x)/sn(x)\right) \cdot \left(P(y)/sn(y)\right)}$$

where x, y mean the word with specific part-of-speech and *sn* means the number of distinct senses for that word listed in WordNet. This is our adaptation of traditional mutual information (MI) in which we take into account the polysemy of the words x and y. In other words our formulation of MI is normalized for the number of senses of x and y. This formulation helps overcome traditional MI's underextraction of collocations that contain high frequency words. For example, our normalized formulation of MI detects the verb *take* as one of the top collocating verbs with the noun *temperature* (as in *The nurse took the patient's temperature*) whereas traditional MI would not detect *take* as a collocate of this noun. (See [14] for details). These

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

word probabilities are extracted from BNC.

All possible pairings of POS-specific words (x and y above) in which the two words appear within a five-word window of each other in our 20 million-words of running text of the BNC are taken as collocation candidates. Using the above algorithm, each x,y ordered pair yields a word association score. Collocations are word pairs that show a sufficiently strong word association between the two words in the pair. Thus, a minimum score threshold is used to select which of the candidate word pairs constitute collocations. This threshold can be lowered or raised to adjust the agent's precision and recall. The collocation knowledge thus extracted from our POS-tagged BNC feeds our browser-based G-Collocator, enabling the agent to detect and highlight collocations that appear in the web pages that the user browses. Figure 2 shows a sample interface with the display of collocations detected by G-Collocator on a specific browsed web page. The detected collocations are listed on a dropdown menu. Each of these listed collocations then links to further examples of the same collocation from BNC and to a highlight option, which triggers the agent to highlight the collocation within the current webpage for the learner's convenience. The check mark to the left of a collocation on the dropdown list indicates the collocations that the user has requested to be highlighted within the web page text.

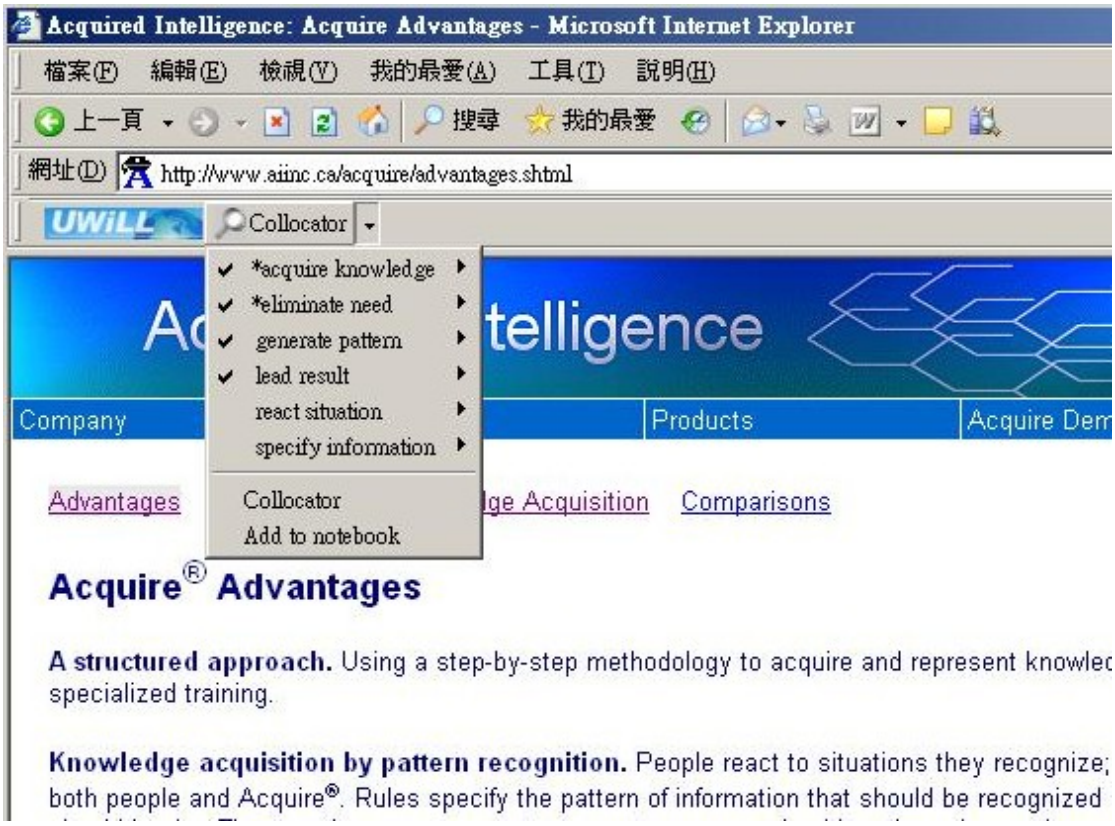

Figure 2

5.3 P-Collocator (PC)

There is reason to believe that in order to master the production of collocations, learners need something more than exposure to positive examples of these collocations. Collocations typically entail two dimensions of knowledge: (1) knowledge that the two words in a collocation are a conventional pairing, such as *heavy rain* or *strong wind*; and (2) knowledge that the collocate (*heavy* in *heavy rain* and *strong* in *strong wind*) is not freely substitutable, that is, knowledge that the collocation *strong wind* can not be paraphrased as *heavy wind* or *big wind*. This second aspect of collocation knowledge is sometimes referred to as non-substitutability. We have anecdotal evidence that the learners who grasp the first dimension of a collocation do not necessarily grasp its second negative dimension as a corollary. Specifically, Wang [9] found in her pretests of her foreign language learner subjects' collocation knowledge that a substantial portion of subjects who were able to supply the correct collocate for a specific focal word in a production task also incorrectly judged the counterpart miscollocation to be acceptable as well. For example, a subject who correctly supplied the verb *spend* as the collocate of *time* (i.e., *spend time*) also incorrectly judged the miscollocation *pay time* to be acceptable as well. The weakness of G-Collocator is that it addresses only the first dimension of collocation knowledge. It conveys to learner that *take medicine* is a collocation whenever this pair is encountered in browsing, but it does not let this learner know that *eat medicine* is not an acceptable alternative of this expression, for example.

On the assumption that learners require both dimensions of collocation knowledge, the motivation for P-Collocator is to add to G-Collocator the second dimension: relevant unacceptable miscollocations associated with detected collocations indicating the non-substitutability that is not apparent from positive examples alone. To do this, we need a learner model, and for this learner model we need an additional knowledge source: knowledge of the miscollocations that the target learner population produces. On the basis of these attested miscollocations, P-Collocator provides personalized or picky collocation detection (hence, the P-). It does this by piggy-backing on G-collocator's results, adding our learner model and a mapping between the learner model of attested miscollocations and pinpoint domain knowledge, that is, the corresponding correct collocations.

The learner model consists of an archive of attested collocation errors found in our 3-million-word corpus of English produced by learners in Taiwan (called English Taiwan Learner Corpus or EnglishTLC). The pinpoint mapping between this learner model and specific target domain knowledge consists of pairings between each of the collocation errors in the

learner model on the one hand and its counterpart correct collocation (or collocations) on the other. Piggy-backing on the collocations that G-Collocator detects, P-Collocator thus is able to determine which of these collocations that have been detected in the current webpage map back to attested miscollocations in the learner model. For example, G-Collocator will detect in a current webpage that *acquire knowledge* constitutes a collocation. P-Collocator can map this collocation onto the learner model and detect that *learn knowledge* and *get knowledge* are attested miscollocations that learners have produced instead of the correct *acquire knowledge*. Next, we describe the design of these main components in this P-Collocator's architecture.

The two main components of P-Collocator knowledge are the learner model and the mapping between standard collocations and their corresponding miscollocations produced by learners. As mentioned above, the learner model is derived from our EnglishTLC learner corpus. We use two methods to extract miscollocations from EnglishTLC. First, since the learner corpus has been partially error-tagged by teachers (See [11]), miscollocations thus tagged serve as one source of the LM. Second, using semi-automatic techniques we bootstrap from these tagged miscollocations to uncover additional, untagged, miscollocations in the learner corpus [6][12].

The second component is the mapping between each of these attested collocations in the LM to the corresponding correct collocations. An example of this would be the miscollocation *pay time* on the one hand and the correct version *spend time* on the other. A portion of these pairings have been provided by hand and designed into regular expression rules that detect and correct learner miscollocations at a 96% precision rate [6]. To supplement these, Wible et al [10] proposed a computational tool called Lexical Assistant which is designed to take as its input attested miscollocations from EnglishTLC and return an acceptable collocation. They hypothesize that the correct alternative to a miscollocate is likely to be found among the synonyms of that miscollocate or among other semantically similar expressions. For example, for the mistaken expression "Did you eat your medicine yet?" the correct counterpart for 'eat' here, that is, 'take', is indeed a synonym of 'eat' in one sense of 'eat' and in one sense of 'take'. In this respect, the very nature of collocation errors suggests that a valuable source for their correction is the synonym set of the wrong word. Lexical Assistant exploits the data structures of WordNet. Since, WordNet encodes other lexical semantic relations in addition to synonymy, we are able to search WordNet not only for the synonyms of the miscollocate but also for its hyponyms and hypernyms as well in order to systematically expand the set of candidate corrections for the miscollocate.

With the LM of attested miscollocations and with our mapping function that provides the correct collocation for these errors, P-Collocator not only detects collocations in the noise of the

unrestricted Web, it also points out to the user that this particular detected collocation is the correct one that should be used instead of a particular common miscollocation often produced by this population of learner. For example, upon detecting and highlighting *acquire knowledge*, P-Collocator also points out that this is the correct version of the common error *learn knowledge*.

The interface for P-Collocator is illustrated in Figure 3. Notice, that since the collocations detected by P-Collocator are a subset of those detected by G-Collocator, we can show the results of both on a single list. The entire set of detected collocations appears on a dropdown list from the toolbar. These are the collocations detected within the current webpage by G-Collocator. The subset of these detected by P-Collocator is indicated on this same list by the addition of an asterisk * (for example, the top two collocations on the dropdown list in Figure 3—*acquire knowledge* and *eliminate need*). By clicking on any of these asterisked collocations, the user can display P-Collocator's matching of this collocation to the incorrect one often used by this population of learners. Figure 3 shows the results of clicking on *acquire knowledge* from the list. This triggers the display to the right of the dropdown list, where the learner can encounter both dimensions of this collocation: *acquire knowledge* is a collocation, and *learn knowledge* is a corresponding attested miscollocation to avoid.

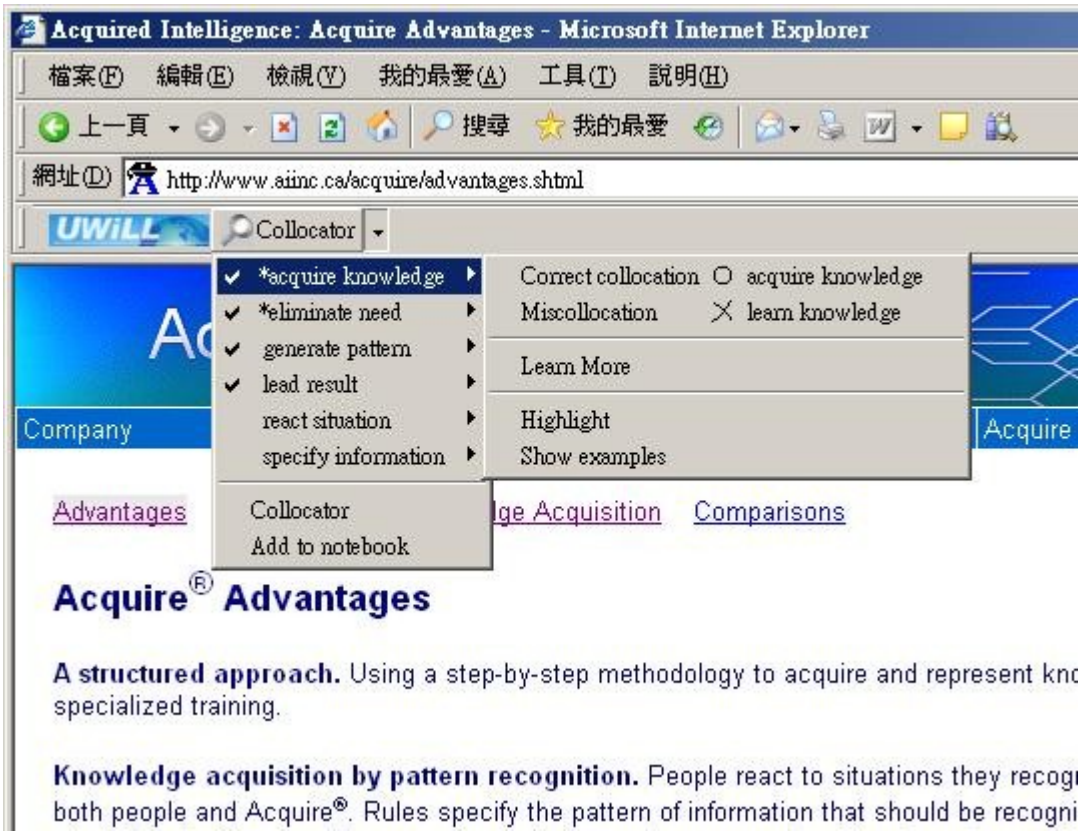

Figure 3

5.4 Notebook Function

The experience of web browsing is notoriously fleeting an ephemeral. In order to allow the collocation input provided by Collocator to take root in the learner's second language competence, there is a need to supplement the exposure that Collocator provides to these collocations that they encounter during browsing on the Web. It is widely acknowledged in the second language vocabulary research that repeated exposure is one of the fundamental requirements that must be met if vocabulary item is to be acquired (See [2] for a review of this literature). To create opportunities for repeated exposure from the fleeting contact with collocations on the Web, we add a notebook function for each learner. The notebook can be displayed on the left of the browser interface (See Figure 4). It allows users to store and organize any of the collocations Collocator highlights or any of the additional example sentences that Collocator provides. It also supports searches for other collocations not encountered during browsing. In addition, as the key to repeated exposure, it offers a "push" function that enables the user to request repeated exposure to a particular collocation over subsequent days.

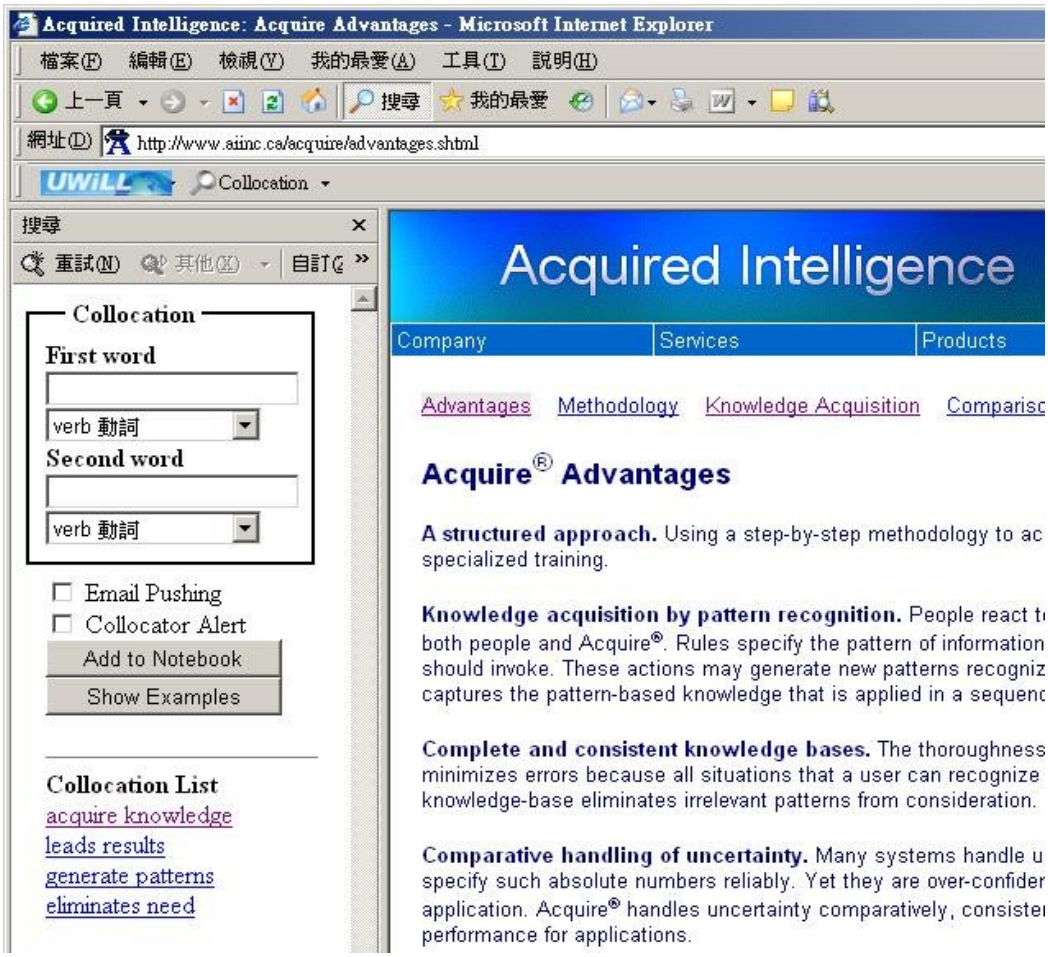

Figure 4.

6 Future Directions

Collocator supports increased personalization of the LM by referring not only to the aggregate learner corpus for miscollocations, but also to an archive of English written by the individual user to detect errors produced by that learner. This is possible because Collocator is incorporated within the architecture of a larger online platform, IWiLL (See [11]), which automatically archives the writings that learners produce on the platform, for example as writing assignments turned in to a teacher on the platform or writings the learner has posted to any of the discussion boards on the platform. One current limitation of the personalization approach is that the small amount of individual learners' written production causes low recall of that learners' collocation problems. With these individual archives of written production currently in place within the system architecture, however, the effectiveness of this personalization of the LM will grow as individual learners' written production accrues over time.

References

1. Thorsten Brants, "TnT-A statistical part-of-speech tagger," in *Proceedings of ANLP-2000*, Seattle, Washington, (2000).
2. Judie Feng-Yi Chien, *A Study of Input and Second Language Lexical Acquisition*, Master Thesis, Department of English, Tamkang University, Taiwan, (2003).
3. Kenneth Church, William Gale, Patrick Hanks and Donald Hindle, "Using statistics in lexical analysis," in *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, edited by Uri Zernik, 115-164. Lawrence Erlbaum Associates, (1991).
4. J. R. Firth, "A Synopsis of Linguistic Theory 1930-1955" in Studies in Linguistic Analysis, Philological Society, Oxford; reprinted in Palmer, F., (ed. 1968), Selected Papers of J.R. Firth, Longman, Harlow, (1957).
5. Stephen Krashen, *Principles and Practice in Second Language Acquisition*, Cambridge: Pergamon Press (1982).
6. Anne Li-Er Liu, *A Corpus-based Lexical Semantic Investigation of Verb-Noun Miscollocations in Taiwan learners'English*, Master Thesis, Department of English, Tamkang University, Taiwan, (2002).
7. Christopher D. Manning and Hinrich Schutze, *Foundations of Statistical Natural Language Processing*, The MIT Press, (1999).
8. R. Schmidt, "Attention" in P. Robinson (ed) *Cognition and Second Language Instruction*,

Cambridge: Cambridge University Press, (2001).

9. Wanju Sonia Wang, *The Effects of Degrees of Explicitness of Automated Feedback on English Learners' Acquisition of Collocations*, Master Thesis. Department of English, Tamkang, Taiwan, (2005).
10. David Wible and Anne Liu, “A Syntax-Lexical Semantics Interface Analysis of Collocation Errors,” *PacSLRF (Pacific Second Language Research Forum)*, October , University of Hawaii, Manoa, Hawaii, (2001).
11. David Wible, 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,” in *Computers and Education* vol. 37, pp. 297-315, (2002).
12. David Wible, Chin-Hwa Kuo, Nai-Lung Tsao, Anne Liu, Hsiu-Ling Lin, “Bootstrapping in a Language Learning Environment,” in *Journal of Computer-Assisted Learning*, vol 19 #1, pp. 90-102, (2003).
13. David Wible, Chin-Hwa Kuo, and Nai-Lung Tsao, “Contextualizing Language Learning in the Digital Wild: Tools and a Framework,” in *Proceedings of IEEE International Conference on Advanced Learning Technologies (ICALT)*, Joensuu, Finland, (2004).
14. David Wible, Chin-Hwa Kuo, and Nai-Lung Tsao, "Improving the Extraction of Collocations with High Frequency Words," in *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, (2004).
15. David Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages189–196, (1995).