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(1/3)

計畫類別： 整合型計畫
計畫編號： NSC93-2524-S-032-005-
執行期間： 93 年 05 月 01 日至 94 年 04 月 30 日
執行單位： 淡江大學英文學系

計畫主持人： 衛友賢

報告類型： 精簡報告
報告附件： 出席國際會議研究心得報告及發表論文
處理方式： 本計畫可公開查詢

中　華　民　國 94 年 4 月 22 日

分散式認知, 電腦運算與隨處的數位語言學習-運用自然語言處理於建置網路英語學習環境(1/3)

語言習得, 分散式認知與學習科技的交集(1/3)

計畫編號：NSC 93-2524-S-032-005
執行期限：九十三年五月一日至九十四年四月三十日

主持人： 衛友賢　淡江大學英文系

Introduction

The purpose of this subproject is to provide the theoretical framework and methodologies for investigating the effects that the digital tools and resources designed in the other two subprojects wield on the language learning of the target users and to supply criteria for the design of these tools, criteria motivated by research in instructed second language acquisition and distributed cognition. The targeted domain of language learning is narrowed to English as a second language and to vocabulary learning. Within this first year of funding, the project has produced substantive results in the construction of such a theoretical framework and in the design requirements that have led to breakthroughs in the design of novel digital tools for ubiquitous language learning. These results have attracted the attention of international scholars and led to concrete international collaborations.

Purpose and Objectives

There are two main purposes of the overall grand project reflected directly in the mission of this individual subproject. (1) To develop a theoretically well-motivated model for browser-based ubiquitous language learning on the web, and (2) To provide an accompanying theoretical framework for investigating the effects of this model on the language learning of the targeted users. In addition to these two purposes shared with the grand project, this individual project has an additional major purpose: (3) To articulate the design criteria for the novel digital tools developed in the other individual projects to support ubiquitous online language learning.

There is a tendency in digital learning research to focus on the development, and dissemination of the technology rather than on carefully investigating the mediating effects that this technology exerts on learning. In fact, at least in the domain of digital language learning, there is no well-articulated research framework or set of frameworks to bring coherence to the development of learning technologies or to the investigation of language learning that is mediated by digital technology. Since learning is supposedly the ultimate desired result of learning technology, the purpose of this three-year project is to develop and implement such a framework for the specific domain of second or foreign language learning. The role of this individual project within this scheme is to provide this framework and the design requirements for the digital tools

To strengthen the focus of the project and increase the chances of creating substantial contributions within this three years, the scope is narrowed to vocabulary learning.

The Framework

Two notions form the center of the approach to digital learning developed in the overall grand project: ubiquitous online learning and distributed cognition. It is a main responsibility of this subproject to articulate these notions and propose a detailed theoretical framework for both implementing them and investigating their effects on language learning. A second major responsibility of this subproject is to provide design criteria for the digital tools developed under the other individual projects. With the notions of distributed cognition and ubiquitous online language learning at the center of this project, we take seriously our mandate to contribute to learning theory. Our progress in the development of these two notions is described in what follows.

Ubiquitous Online Language Learning

This notion of ubiquitous online language learning is a novel conceptualization for implementing language learning support in noisy digital environments. It contrasts sharply with both current digital learning approaches as well as traditional classroom learning approaches. The conceit of our approach is that it creates a natural fit between the structure and content of the World Wide Web as a source of language input for language learners on the one hand and second language acquisition theory which gives a primary role to target language input for learners within authentic communicative contexts that fit the learners' own purposes and interests on the other.

In the first year of the project we have succeeded in disseminating this novel approach to digital language learning within the international research community. Some examples follow.

One of our two full papers presented at IEEE's ICALT2004, entitled "Contextualized Language Learning in the Digital Wild: Tools and a Framework," was a

presentation of the philosophy of ubiquitous online language learning along with a description of two novel tools we developed in cooperation with the second individual project (Kuo) which implement this philosophy: Collocator and Word Spider.

At EuroCALL2004 (Vienna) we presented a full paper on a novel design for learner corpora which we call Dynamic Event-Driven (DyED) Learner Corpora, one which we have already implemented in our 3-million word corpus of Taiwan learners' English (EnglishTLC). The design captures not only the texts written by these learners in English, but the contexts of the writing. In other words, while current learner corpora a taken as collection of texts, the DyED learner corpus design takes these texts as writing events that occurred in situated contexts. The corpus indexes the learners' texts to the assignments which evoked the texts, to all of the teacher's feedback on each text, to the revisions that each learner made in response to these instances of teacher feedback, and so on. Thus one piece of student writing in our current version of EnglishTLC is indexed to an entire mesh of related digital events that may exert influences on this writing. Thus the DyED corpus design supports a much richer array of research on student language learning and student writing than other current learner corpora.

This novel learner corpus design has already begun to have its effects on the international research community. Professor Sylviane Granger, former director of the Centre for English Corpus Research at the University of Louvain, Belgium, and the current director of that University's Language Institute (10,000 language students and 70 instructors) has asked to collaborate with us in implementing our system in their Language Institute as the main language learning platform. The ubiquitous tools will be an integral part of the platform we offer, and their Language Institute's use of the tools will provide valuable data on the tools effects on learning. Our research group led a visit to Columbia University, Columbia Teachers College (TC) in June 2004 for exchanges with Professor Jo Anne Kleifgen and her colleagues and graduate students on topics of digital language learning. The trip was a remarkable success, and we are now finalizing plans for a two-day research roundtable on digital language learning at Columbia University for the summer of 2005 to include more senior Columbia researchers and more researchers from our group in Taiwan.

Distributed Cognition

We have taken the notion of distributed cognition as central to our way of conceptualizing the relationship between the language learner and the digital tools that learner uses. Our contribution is to apply this construct to the framing of investigations of digital learning. The role for distributed cognition within a research framework for digital language learning has been described and motivated in detail in Wible (2005) "Distributed Cognition, Learning Events, and Research on Digital Language Learning" (submitted to *Computers and Education*, Nov 2004, ms 54 pp.).

Distributed cognition plays two sorts of roles in our project: (1) as a construct in the research framework that we are developing, and (2) as a construct in the empirical investigations we are currently designing and conducting.

Distributed Cognition in our Research Framework

A central insight from distributed cognition is that the way in which knowledge is stored and processed by humans is determined fundamentally by the details of their environment and by the demands of the tasks which require this knowledge. We propose that this insight has basic implications for research and design in digital language learning. Consider the example of an online dictionary look-up function that accompany online reading and allow the user to conveniently look up any word encountered while browsing the web. Providing learners with access to such a function will not necessarily yield the sort of learning or assistance that the tool is intended to provide. This is because the nature of the situated lexical competence needed by a learner for a variety of contextualized language tasks such as reading for global comprehension or skimming a text for information could be far removed in nature from the lexical information as represented to the learner in an online dictionary. It is precisely such distinctions and assumptions provided by the view of distributed cognition which should be part of a framework for productive CALL research and design, distinctions, for example, between lexical information outside the head on the one hand and various situated lexical competences required of language users on the other.

Congruent with these guiding principles from distributed cognition, our choice of the particular areas of lexical competence we focus upon and the tools we design to provide learners assistance with these competences is carefully motivated by the nature of the language competence and the potential for enhancing them with digital assistance. The collocations we treat below are an example where our tools provide the supplement to their language input that is precisely needed to grasp the collocations. It is also why we shun simply providing dictionary look-up functions for online reading, though this sort of tool would be easy for our teams to provide and would be more than welcome by the learners using our platforms.

Distributed Cognition in our Empirical Investigations

One of our empirical investigations will be reported at EuroCALL 2005 in Wible, Kuo, Wang, and Tsao (2005). In that investigation we tracked the learners' use of and attentiveness to automated feedback on collocation errors they made in both translation exercises and essay writing. Moreover, to investigate the effects of particular types of feedback on collocation errors, three groups received

feedback of different levels of explicitness. The results show that the explicitness of digital feedback as well as the attentiveness of subjects to this feedback had statistically significant differential effects on the learners' improvements in their use of the targeted collocations over a one month period.

The next stage of our tool development and our research integrates it with ubiquitous online learning. Specifically we link the automatic feedback that learners get on their writing to the sorts of help provided to these learners when the browse the web subsequent to receiving this feedback. The prototype tools for this have already been completed. The design works as follows.

Once a learner produces in their own online writing an error in the use of one of the targeted collocations, the system not only provides feedback and positive examples of the correct collocation (for example correcting *pay time* with *spend time* and showing corpus examples of this); the system also records this event in the individual profile of this learner and tracks their web browsing. When the learner browses any webpage that contains the correct collocation, the system offers to show this to the learner by highlighting the collocation in the text. Records of the learners' attentiveness to this feedback are kept as well, for future research.

Domain Knowledge Results

An important feature of the research project is a clearly defined focus on a narrow scope of the learning domain. Specifically, our work is limited to the learning of lexical knowledge (vocabulary acquisition). Within this domain, a crucial research niche is how digital tools (specifically computational linguistic techniques) can be applied in novel ways to extract the sorts of linguistic knowledge that is intended for the learners. In this case, it is how to extract lexical knowledge. On this front, we have this year made novel contributions in computational lexicography. In fact, we could say that we have seen a breakthrough in our work which sets it apart from other work being done internationally within the same domain. This work is a close collaboration between this individual project and the NL2P individual project—a combination of linguistic and pedagogical expertise on the one hand and computational techniques on the other. What follows is a brief progress report on this work.

Learnability and Collocations

Two of the digital tools developed in Individual Project 2 (Kuo) are designed to detect collocations. The Collocator in particular is designed to do this in noisy unrestricted environments where learners browse freely. Here we motivate the tool from the perspective of language learning and learnability.

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 and why learners require vast amounts of accumulated experience with the target language to acquire collocations. By detecting collocations in real time in noisy environments, the Collocator tool overcomes this obstacle and distills into one reading experience the exposure to a collocation that would usually require hundreds of hours of input from the target language.

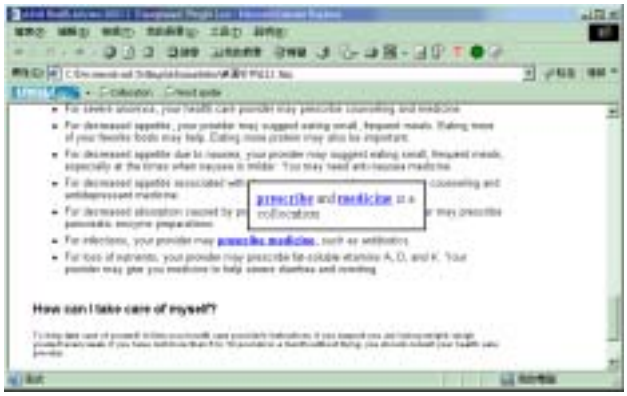

Figure 1. Collocator

Lexical Clustering

There is a substantial amount of research on automatically extracting lexical knowledge from large corpora using stochastic methods. Some of this work is particularly focused upon automatically identifying collocations (such as a *warm welcome*, *prescribe medicine*). We have developed a tool called Collocation Explorer that has been judged by English teachers at Taipei First Girls Senior High (TFG) as the best collocation resource for learners they have seen (whether

digital or book form). This tool now serves as an everyday part of the IWiLL learning activities, with regular 'collocation challenges' designed by core English teachers and posted on the IWiLL front page for all the learners in Taiwan to participate.

An extension of Collocation Explorer has now been developed called Collocator (motivated in the previous section) which can support ubiquitous online collocation learning. It is a browser-based version accessible from toolbar which detects collocations within any webpage the user happens to be browsing.

Our lexical research has extended beyond collocations this year. Collocations are typically word pairings (though some may consist of 3 words). A less well-defined concept of word combinations has become popular in second language teaching recently called lexical chunks. Lexical chunks are combinations of words regularly used together, such as '…with respect to…' or '…in the event of…' These are extremely important components of a proficient English user's lexical competence. They also represent a wide-open research niche which we have identified and are currently working within. While numerous books and articles now urge language teachers to teach lexical chunks, there are no resources providing these chunks. There are no reference books of 'English lexical chunks' (though there are numerous collocation dictionaries). There are also no computational tools that extract chunks (though there are many that extract collocations).

We have targeted this niche then for investigation and aim to provide world-class research and system implementation. In the first year, we have both identified this niche and already developed a prototype of an automatic lexical chunk extractor. It is based on greedy algorithm we have designed which identifies strings of words of any length (including strings of non-contiguous words) that constitute lexical chunks worth learning. For example, given the word 'fact' as input and no other guidance, our current prototype automatically discovers the chunk: 'despite the fact that' as well as collocations like 'In fact' among others. It does this by running our algorithm on 20 million words of the British National Corpus. Given the target word 'point' as input, it detects chunks such as 'from the point of view of…' We are currently refining the algorithm to improve precision and recall and designing an interface for learners. The current interface is for internal research use only.)

This is the only such computational tool of its kind that we are aware of. Others that most closely resemble our chunk detector suffer from having a fixed size of word groups. Kilgariff's WordSketch, for example, provides exactly 3 words including the target word. Moreover, the intended users of WordSketch are professional lexicographers, so the interface is opaque for language learners and difficult to grasp. Other collocation tools find only word pairs. Ours is the only one that is completely flexible with respect to the length of chunk it can detect (both …*in fact*… as well as …*despite the fact that*… are detected by the same algorithm. Moreover, the intended users for our tool are language learners, not professional lexicographers.

The next step after fine-tuning our clustering algorithm will be to implement it as a ubiquitous tool, available on our browser toolbar. There it will be able to detect the occurrence of lexical clusters in the texts that the user freely browses on the web, providing the authentic context chosen by the learner, raising their level of investment in the learning experience. This is in stark contrast to the approach of preparing lessons which may or may not interest the learner and which users may or may not access.

Future Directions

Currently we are designing focused tracking studies that follow our users ubiquitously during their web browsing and record relevant learning events they engage in, such as accessing our Collocator annotations. In order to investigate the effects of the particular design or content of the ubiquitous tools, we are creating minimally distinct variations of these tools to determine whether these variations lead to differential effects on the users' language learning.

The collocation and lexical clustering tools make it possible for us to gather rich repositories of insights into individual words, for example to investigate the vocabulary words listed on the MOE English Vocabulary reference lists and extract a rich array of properties of these words that can not be found in dictionaries. We are beginning now to assemble an enriched, annotated version of the MOE vocabulary lists to make available online. They will allow the users to browse the vocabulary list, click on any word and display the important collocations and lexical chunks that this word occurs in. This knowledge will also be implemented in a ubiquitous browsing tool that detects lexical chunks in the texts of the webpages that the users freely browse.

Selected Bibliography

Wible, David, Kuo, Chin-Hwa, and Tsao, Nai-Lung, "Contextualizing Language Learning in the Digital Wild: Tools and a Framework" IEEE ICALT 2004, Joensuu Finland.

Wible, David, Kuo, Chin-Hwa and Tsao, Nai-Lung, "Towards an Ontology of Digital Learning Events for CALL," EuroCALL 2004, University of Vienna, Austria.

Wible, David, Kuo, Chin-Hwa and Tsao, Nai-Lung, "Improving the Extraction of Collocations with High Frequency Words," *International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, May 26-28, 2004

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

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