Chapter 2

First Person Singular: A Digital Library Collection that Helps Second Language Learners Express Themselves

Shaoqun Wu
University of Waikato, New Zealand

Ian H. Witten
University of Waikato, New Zealand

ABSTRACT
We use digital library technology to help language learners express themselves by capitalizing on the human-generated text available on the Web. From a massive collection of n-grams and their occurrence frequencies we extract sequences that begin with the word “I”, sequences that begin a question, and sequences containing statistically significant collocations. These are preprocessed, filtered, and organized as a digital library collection using the Greenstone software. Users can search the collection to see how particular words are typically used and browse by syntactic class. The digital library is richly interconnected to other resources. It includes links to external vocabularies and thesauri so that users can retrieve words related to any term of interest, and links the collection to the web by locating sample sentences containing these patterns and presenting them to the user. We have conducted an evaluation of how useful the system is in helping students, and the impact it has on their writing. Finally, language activities generated from the digital library content have been designed to help learners master important emotion related vocabulary and expressions. We predict that the application of digital library technology to assist language students will revolutionize second language learning.

INTRODUCTION
Everybody wants to talk about themselves: their thoughts and feelings, what they have been doing and what they plan to do. In other words, we all aspire to become expert in the first person singular. But in a foreign language, it is not easy. Language learners often complain that they cannot express what they think, feel and do. You might answer a simple question like “How are you today?” factually (“My head aches”), perfunctorily (“OK”), or provocatively (“I’m feeling sexy”). But students
find it hard to go beyond simple statements and talk about their feelings at greater depth. And the same applies to all forms of self-expression.

Part of the reason is that learners have not experienced enough of the language to express themselves in the first person in ways that sound natural. As Moskowitz (1978) notes, curricular material tends to focus on facts and everyday transactions, only rarely touching on vocabulary that is appropriate for communicating more subjective aspects of everyday life. To help remedy this she advocates integrating a humanistic approach to language teaching with a planned curriculum to promote self-actualization and self-esteem, so that students can express themselves meaningfully in the first person.

To be able to talk fluently about themselves, learners must command appropriate linguistic resources. This paper describes how to identify short sequences starting with (or, in some cases, containing) the word “I” and use them to help learners acquire important “I-vocabulary” and “I-expressions.” Fluency does not blossom from a comprehensive lexicon of difficult words, nor even from familiarity with the most common ones. Instead, it requires an internalized repertoire of phrases and expressions composed of words used in everyday life (Lewis, 1993). Consequently our digital library focuses on the most commonly used English words and their associated expressions.

How can ordinary, everyday language be captured? Our approach is to capitalize on the text on the World-Wide Web, in particular the vast set of n-grams from the Web that Google has made available. Only digital library technology can provide searching and browsing functions for such a massive body of text. Our system is based on the Greenstone software (Bainbridge et al., 2004). We have built a collection called “First Person Singular” that allows learners (and teachers) to locate phrases associated with a particular word, as well as synonyms, antonyms, and collocations. The digital library enables sentences containing these patterns to be retrieved from the Web and presented to the user as examples. We have conducted an evaluation with actual language students, and the results show the potential usefulness of the system in helping students correct grammar errors, generate text and expand text.

In this paper we first examine the n-grams Google has supplied and explain how to extract a subset that is useful for language learning. We then describe the design and implementation of the First Person Singular digital library collection: how it is built and the searching and browsing facilities it includes. Next we show how results obtained from the collection can be augmented by retrieving related material from the Web and the British National Corpus (BNC). Then we describe the findings from an evaluation with actual students.

We round out the paper by describing some language activities that we have designed to help students master important vocabulary and expressions. Although these have not been evaluated formally, they point the way to an exciting future. We believe that digital libraries in general—not just the First Person Singular collection described here—have the potential to revolutionize the area of second language learning by providing unlimited volumes of practice exercises that are generated automatically, directly from a library’s contents. This general strategy will allow any digital library collection to be used as a basis for language learning exercises.

**N-GRAMS FOR LANGUAGE LEARNING**

Our starting point for this digital library is a corpus of word n-grams in English, ranging from unigrams or single words to 5-grams or sequences of five consecutive words, along with their frequency counts. These were generated by Google from approximately one trillion word tokens of text on publicly accessible web pages: a staggeringingly large body of natural English. N-grams
Related Content

Changing Scenario From Information Management to Knowledge Management
[www.igi-global.com/chapter/changing-scenario-from-information-management-to-knowledge-management/188030?camid=4v1a](www.igi-global.com/chapter/changing-scenario-from-information-management-to-knowledge-management/188030?camid=4v1a)

Personal Digital Libraries
[www.igi-global.com/chapter/personal-digital-libraries/19868?camid=4v1a](www.igi-global.com/chapter/personal-digital-libraries/19868?camid=4v1a)

Automatic Text Summarization in Digital Libraries
[www.igi-global.com/chapter/automatic-text-summarization-digital-libraries/19879?camid=4v1a](www.igi-global.com/chapter/automatic-text-summarization-digital-libraries/19879?camid=4v1a)

User Profiles for Personalizing Digital Libraries
[www.igi-global.com/chapter/user-profiles-personalizing-digital-libraries/19878?camid=4v1a](www.igi-global.com/chapter/user-profiles-personalizing-digital-libraries/19878?camid=4v1a)