Jonathan J. Webster & Chunyan Ning, City University of Hong Kong

WWW Bilingual Chinese-English Language Dictionary Database

Abstract

The application described in this paper is intended to facilitate web access to lexical information for Chinese and English. A Visual Basic application connects front-end HTML forms with a back-end object-oriented database implemented using LPA Prolog++. The three main components of the application are (1) the back-end database for storing lexical information; (2) a Web Client application for front-end query input; (3) an Agent application to handle the exchange between the client's queries and the database. The access strategies being incorporated into this application are based on natural classes of lexis defined according to thematic, taxonomic and semantic relations.

Accessing dictionary data on the internet

The on-line bilingual Chinese/English language dictionary we are developing will provide web access to lexical information for Chinese and English. Our aim is to build an on-line bilingual dictionary data retrieval tool for language learners and language professionals which duplicates the human attempt at defining words, i.e. defining word meaning based on associations or relations with surrounding words and phrases. The back-end dictionary database is intended to provide comprehensive coverage of Chinese and English words. Users will be able to (a) retrieve dictionary data for Chinese/English words/phrases as with a conventional dictionary; and (b) retrieve entries matching syntactic and/or semantic criteria given by the user. The three main components of the application are (1) the back-end database for storing lexical information; (2) a Web Client application for front-end query input; (3) an Agent application to handle the exchange between the client's queries and the database.

How will our WWW Bilingual Chinese-English language dictionary database compare with other attempts at providing internet access to lexical/dictionary data? There are, for example, several 'Webster Servers' available on the internet. Examples (as reported by Steeve McCauley, http://www.eps.mcgill. ca/~steeve) include webster.cs. indiana.edu (restricted access), citi.umich. edu (open), webster.cs.

mcgill.ca (restricted), webster.eps.mcgill.ca (experimental). McCauley's 'Windows Webster Client' is an example of software which facilitates downloading dictionary data from such servers. Typical dictionary data is generally retrievable much as one would expect to find in an ordinary dictionary.

In a recent paper, Webster (1995) reports on development of an application in which HTML forms serve as the front-end to a lexical database. Lexical information and data retrieval strategies are based on the *Longman Language Activator* (LLA). A Visual Basic CGI application connects a front-end HTML form with the back-end relational database implemented in MS Access. The LLA was chosen for its unique organization of dictionary information which is intended to make it easier for the user to find the right word or phrase for a particular context. The LLA adopts three access strategies: first and foremost by concept using the Key Words, second by entry word, and third by what the LLA calls 'access maps'. The home page for the application is, in fact, several forms, the first and top-most consisting of the field into which the user enters a word/phrase to search for and a submit button for transmitting a URL request to the web server. The URL request is a VB CGI executable program which queries the Access database, and returns the word list information for the search word. The design of the program is modeled after examples of VB/Access CGI programming provided by R. Denny, the designer of WinHTTPD and the Windows CGI, and also examples discussed in Heslop and Budnick (1995). Basic CGI initializing operations are handled by Denny's CGI.BAS module. This application was experimental only. No attempt was made to enter all the information contained in the LLA. The primary objective was to demonstrate the potential of rendering a particular approach to lexical information retrieval in the form of a hypermedia presentation for easy web access. The application was implemented in Windows 95 using 32-bit VB 4.0 and Access 7. The web server was O'Reilly's 32-bit Website 1.0.

For this application, instead of opting for the proven but less efficient VB-CGI approach, we are experimenting with the Wayfarer QuickServer SDK to facilitate client-server operations over the internet. A particularly interesting feature of the QuickServer is that it permits the use of VB (other languages such as Java, C++ are supported as well) to create Web Client applications. Using the plug-in feature of Netscape Navigator 2.0 and Wayfarer's own Web Extension Manager, a QuickServer client application may be exposed within the browser. The user submits queries from the client application embedded in the browser. In other words, VB forms appear just like any other web pages. The client then connects with the server and binds with the QuickServer Agent. The Agent, also a VB

application, handles all communication with the dictionary database developed in LPA Prolog++. Prolog++ is a programming language which combines Prolog's declarative approach with object-oriented programming. An object-oriented approach is ideally suited to maintaining the dictionary database of natural classes of lexical items described above. The link between Prolog++ and VB is made possible by means of a dynamic link library written in C.

Natural classes of objects

Whether speaking in lay or linguistic terms, however we choose to characterise a lexical item is far from the totality of its meaning in actual use. This is due to the large quantity of factors, both linguistic and extra-linguistic, which together realise the meaning of the word in context. Language itself is a system evolving out of speech encounters in which "people create meaning by exchanging symbols in shared context of situation" (Halliday 1984:11). Semantically-based verbal relationships exist along a vertical or semiotic dimension extending from context of situation through text to clause. Words, like other semiotic units, "obey the Gestalt principle of having overall properties transcending the mere sum of their parts, and functioning in their contexts as integrated wholes" (Garvin 1985:57). For example, the dictionary definition of the word *door* can only begin to 'scratch' (Chomsky 1993) a tiny surface of its meaning. One sense of the word appears in the sentence *John is knocking at the door*, while a second sense is realised in *John is walking through the door.* The meaning of the word *door* is naturally defined by its association with the activity of either 'knocking at' or 'walking through'. To duplicate the human lexicon for computational purposes one must endeavour to duplicate the human attempt at defining words, i.e. defining word meaning based on associations or relations with surrounding words and phrases.

Given that the size of the lexicon in language A is S containing n number of lexical items $< L_1, L_2, ... L_n>$, the meaning of the lexical item L_1 is defined as follows:

1. The meanings of L_1 ::= all its possible relations with $< L_2, ... L_n >$
2. $Meaning_1$ of L_1 ::= its possible relations with a natural class NCL_1 consisting of members from $< L_2, ... L_n >$
3. $Meaning_2$ of L_1 ::= its possible relations with a natural class NCL_2 consisting of members from $< L_2, ... L_n >$

4. Meaning$_N$ of L_1 ::= its possible relations with a natural class NCL_N consisting of members from < L_2, ... L_n >

There are natural constraints on what items (including any that might subsequently be coined) can join a natural class in entering into a relation with L_1 so that the members of the set < L_2, ... L_n > can be fully specified. For example, whatever new nouns enter the V-N relation with the verb *sell* will be a member of the natural class otherwise containing [*books, houses, city, idea, ...*].

Among the relations in (1–4), the following three are included in the design of our database:

a. thematic relations,
b. taxonomic relations (subordination and superordination), and
c. synonym/antonym relations.

Thematic relations can be established in an SVO language like English or Chinese by the following three rules:

(5) i. VP -- [L_i, N], if L_i is a verb;
ii. VP -- [V, L_i], if L_i is a noun;
iii. S -- [N, L_i], if L_i is a VP.

The output of these rules are three types of relations rendering the natural classes illustrated in (6–8) below:

(6) i. open <door, window, car, box, eye, ...>
ii. open <bank-account, ...>
iii. open <bank, restaurant, school...>
(7) i. <knock-at, paint, kick, ...> door
ii. <walk-through, enter, ...> door
(8) i. <bank, school, restaurant,...> open
ii. <door, window, eye, mouth...> open

Each line in (6–8) defines a particular sense of a word in use. When applied bilingually, there appears a one-to-one correspondence between the contextual meaning of a word in one language and the corresponding contextual meaning in another:

(9)	**Chinese**	**English**
	kai <qiche, motuo, houche...>	drive <car, train...>
	kai <deng, dianshi, jiqi,...>	turn on <light, TV ...>
	....	....

The taxonomic relations of subordination and superordination are also listed as exhaustively as possible according to the following rule:

(10) L_i <L_j, ... L_n>, where L's are all nouns.

For example, a noun like *car* may hold a subordinate relation to a fixed set of other nouns available in the lexicon:

(11) car <vehicle, transportation, commodities, machinery, ...>

Whereas a noun like *transportation* may hold a superordinate relation to a fixed set of nouns:

(12) transportation <car, boat, bike, ...>

Synonyms and antonyms of a word with an identifiable contextual meaning form additional classes.

Thus, the human lexicon is an inventory of words as objects. Corresponding to each word are natural classes of (a) other words or objects in thematic relations, (b) other words in taxonomic relations, and (c) other words in synonym/antonym relations.

Dictionary Data

The information being included in the database for each entry includes the following: simplified and traditional scripts, romanization (both pinyin and Cantonese). The entry data for 打 is illustrated below:

Name	Value
Simplified / 簡化	打
Traditional/ 繁體	打
Pinyin/ 拼音	da3
Cantonese/ 粵語	daa5

Following an object-oriented approach, each word is treated as an object or class. As illustrated below, 打 is declared as a class which subclasses from the super class lexeme_Chinese. It therefore inherits certain properties from its super class, lexeme_Chinese.

```
class lexeme_Chinese.
attribute simplified, traditional, pinyin, cantonese.
...
end lexeme_Chinese.

class '打'.
inherit lexeme_Chinese.
simplified('打').
...
end '打'.
```

Information about natural classes defined by their thematic relation with the entry word is essential to determining semantic variation. For example, three potential senses of the entry word 打 are realized in the context of phrases initiated by the entry word. The first has to do with hunting animals, e.g. 打（虎/狼/野鷄/野豬/...）; the second with catching fish, e.g. 打魚. The third meaning involves playing games or sports:

打球	play ball
打高爾夫（球）	play golf
打麻將	play mahjong

The three senses translate into three different words in English, i.e. *hunt*, *fish*, and *play*. An equivalent translation in English is considered to have been found if there is a successful though not necessarily perfect match in terms of the class of objects thematically related.

Also accessible for each sense must be the relevant syntactic and lexico-semantic features, such as syntactic category, process type and complementation type. The notion of process type corresponds to Halliday's system of transitivity which specifies various types of processes, e.g. material, mental, relational, behavioural, etc. Comple-mentation type refers to whether the verb is intransitive, monotransitive, ditransitive, etc. Again, in the case of 打, the following information must be entered into the database:

Name	Value
Syntactic Category	verb
Process Type	material
Complementation type	monotransitive

Similar to Tang's (1995) object-oriented Chinese lexicon, verbs in our database inherit attributes depending on who their parents are in a class hierarchy consisting of super classes designating various process and complementation types.

While for the entry for "籃球/basket ball" (referring to the object not the game), the following syntactic information would be stored for subsequent retrieval:

Syntactic Category	noun
Countable	yes
Classifier	個
Animate	no

Also included are those natural classes defined by taxonomic relations, for example for 天氣/weather,

Superordinate:	氣候	climate
Subordinate:	下雪	snow
	暴風雨	rain

Besides synonym/antonym relations, other natural classes defined by such relations as cause-result, material-product, possessor possessee, etc, may also be included. For example, again for 天氣/weather, a possible natural class defined by cause-result includes:

Cause-Result	感冒/疾病/健康/不舒服.....
	flu/disease/health/discomfort.....

References

Chomsky, Noam. 1993. *Language and Thought*. London, Moyer Bell.

Garvin, Paul. 1985. An Empiricist Epistemology for Linguistics. *The Eleventh LACUS Forum 1984*. Hornbeam Press, pp. 331–351.

Halliday, M.A.K. 1984. On the Ineffability of Grammatical Categories. *The Tenth LACUS Forum 1983*. Hornbeam Press, pp. 3–18.

Halliday, M.A.K. 1985. *An Introduction to Functional Grammar.* London, Edward Arnold.

Tang, Nora. 1995. Developing a Chinese lexicon using an object-oriented approach for use with the Chinese Lexical Functional Grammar (C-LFG) Parser. BALIS Honours Project. City University of Hong Kong.

Webster, J.J. 1995. *Proceedings of the The 10th Pacific Asia Conference on Language Information and Computation.* Language Information Science Research Centre, City University of Hong Kong.