Computer Assisted Language Learning
2000, Vol. 13, No. 2, pp. 167–182

0958–8221/00/1302–0167$15.00

An ESL Writer's Collocational Aid

C.-C. Shei and Helen Pain
Division of Informatics, University of Edinburgh

ABSTRACT

Collocation is one of the most difficult aspects in second language learning, but has been largely neglected by researchers and practitioners. A questionnaire survey shows advanced Chinese learners' collocational ability in English to be significantly inferior to that of native speakers. Our research attempts to correct this problem by developing an on-line correcting program which is able to detect some collocational errors in the learner's English writing and offer examples of standard collocations from a large corpus for reference. The system is based on two kinds of corpora: a learner corpus which is used for the study of known collocational errors, and a reference corpus which is used to extract standard English collocations. The system also makes use of a Dictionary of Synonyms derived from WordNet to discover the potential collocational errors in learners' input, as well as a Paraphrase Database gathered from the learners themselves to help diagnose un-collocational learner phrases. Altogether, it is hoped that the result of this research has not only produced a usable on-line collocational aid, but also demonstrates a simple and efficient way of using learner corpora and reference corpora to support CALL software design.

1. INTRODUCTION

This article presents a new paradigm for teaching English collocations to second language learners. Like an idiom, a collocation is a group of words which frequently appear together in texts. An idiom is known as a phrase which acquires a meaning different from the sum of its individual parts; for example, *kick the bucket* has nothing to do with bucket kicking. A collocation is, on the other hand, the habitual co-occurrence of two or more words whose meaning can be inferred from the parts, but will become less acceptable when one of the elements is replaced by a similar word. For example, *achieve a level* is collocational while *achieve a point* is not.[1] The place of idioms in language teaching is long established but that of collocations is not. This article

1. The British National Corpus has 60 instances of *achieve* collocated with *level* to its right within two words distance, while no instance of *achieve* collocated with *point* was found.

Correspondence: Chi-Chiang Shei, Division of Informatics, University of Edinburgh, 80 South Bridge, Edinburgh EH1 1HN, Scotland. Tel: (0131) 6502725. E-mail: shei@dai.ed.ac.uk.

Manuscript submitted: September, 1999.
Accepted for publication: December, 1999.

reports on a computerized English collocation teaching system currently being developed.

One important aspect of our system is its heavy reliance on language corpora at its development as well as implementation stage. A native speakers' corpus is used for the dual purpose of building a standard English collocation library and providing authentic examples of usage. Also a language learners' corpus is collected as the basis for an error library. This methodology is in tune with dominant concepts in language teaching such as authentic language use and learner needs.

This article is organized into six parts: (1) collocation and language learning, which reviews the literature about the role collocation plays in language acquisition; (2) collocation analysis, which reviews some of the works in English collocation analysis; (3) collocation extraction, which introduces our method for extracting collocations from corpora; (4) an empirical study, which compares native speakers' intuition with the result of corpus processing; (5) learner corpus, which discusses the use of learner corpus in the current approach to language teaching; (6) system design, in which the architecture of the Corpus-Based Collocation Tutor is introduced.

2. COLLOCATION AND LANGUAGE LEARNING

Though the role that collocation plays in language acquisition is an important topic, very few systematic studies can be found that address this issue. One recent study is by Zhang (1993), in which a series of experiments were conducted to explore the relationship between the knowledge of collocation and proficiency in writing. It was found that more proficient second language writers use significantly more collocations, more accurately and in more variety than less proficient learners. Gitsaki (1996) further identifies some factors which could influence the development of collocational ability during language acquisition, for example frequency in the input, complexity of the collocations, degree of L1–L2 difference, and the order of collocational parts (e.g., Prep Noun was found to be more difficult than Noun Prep collocations). More research is needed, however, to strengthen and elaborate on these conclusions.

Despite the lack of empirical studies, it is nevertheless generally agreed that collocational knowledge is one of the things which contribute to the difference between native speakers and second language learners. For example, Kjellmer (1991) believes that 'automation of collocations' helps native speakers to utter

sentences more fluently. Language learners, being deficient in this automation, constantly have to create structures and are thus less fluent when speaking. Similarly, Aston (1995) notes that the use of a large amount of prefabricated items speeds language processing in comprehension and production alike, and thus creates native-like fluency.

3. COLLOCATIONAL ANALYSIS

This section briefly reviews some of the research on analysis of English collocations, revealing the interesting nature of collocation in general. Cruse (1986) points out that, for a collocation, each lexical constituent is also a semantic constituent (as opposed to an idiom, where all lexical constituents form a semantic unit). However, Cruse also notes that there is some sort of cohesion among the constituents of a collocation (as opposed to free combinations of words). What is interesting, then, is the nature and types of cohesion for different collocations. So far, to our knowledge, there is no comprehensive discussion on this topic. Some interesting observations are available, however. For example, Cruse observes that it is the notion of 'consumption' which constrains the collocation of *heavy* with *drinker, smoker, drug-user*, etc.

A more general analysis of collocation is provided by Stubbs' (1995) explanation of 'semantic prosody', which approximately means the good or bad connotation a combination of words gives to a reader. For example, the word *cause* is more frequently associated with words of unpleasant meaning, like *cause problems*, whereas the word *provide* usually has positive semantic prosody, like *provide food*. Apparently a second language learner needs to know the mechanism underlying the operation of semantic prosody, though not necessarily the term itself. It is not uncommon for a learner to say *cause happiness*, which of course violates the semantic prosody condition.

Kjellmer (1991) discusses collocational behaviour of parts of speech categories on the basis of corpus analysis. He finds that singular nouns and verb base forms have a higher collocational tendency while adjectives and adverbs are not collocational. This is partially supported by our observation so far while doing the collocation extraction task. In investigating the V–N collocations, it was found that verb base forms almost always have higher collocating ability than other verb forms, like past tense or third person singular (see Table 1 for some evidence). Incidentally, Gitsaki (1996) also found the verb–object type collocation to be the most difficult for second language

Table 1. Partial Frequency List of Verb Collocates for *Confidence*.

Word	Frequency
give	100
gain	66
restore	59
gave	54
lost	54
gives	51
build	47
giving	43
lose	41
gained	38
inspire	33
develop	29
boost	29
undermine	28
maintain	25
say	24
increase	24
expressed	22
said	22
got	22
lack	21
get	17
lacked	16
regain	16
giving	15
increasing	14
retain	13
win	13
need	12
speak	11
find	11
undermined	11
destroy	10
boosted	10
bolster	10

(Frequency under 10 omitted)

learners. This means that it makes sense for us to focus on verb-related collocations initially when prototyping a collocation tutoring system, and to use the verb base forms for analysis while developing the system database.

Other types of collocations should not be neglected, however. For example, Johansson (1993) draws all examples of adverb–adjective pairs from the LOB corpus and makes some interesting classifications:

Degree and extent: *vitally important*
Emphasis: *definitely helpful*
Manner: *enjoyably articulate*
Time: *frequently inaccurate*
Space: *universally unpopular*
Viewpoint and respect: *theoretically possible*
Evaluation of truth: *apparently blameless*
Basic and typical qualities: *typically British*
Value judgement: *absurdly long*

Altenberg (1991) further explores 'amplifier collocations' in English which also involve adverb–adjective collocations. A significant finding in this work is based on the distinction between 'maximizers' and 'boosters', both of which are adverbs which collocate with adjectives and amplify them in different ways. Maximizers (*completely, absolutely, thoroughly*, etc.) modify adjectives which cannot be graded (e.g., *impossible, wrong*), whereas boosters (*extremely, awfully, deeply*, etc.) typically modify fully gradable adjectives (e.g., *big, bad, exhausted*). This explains why *extremely right* in an ESL learner's writing sounds unnatural (wrong match of a booster with a non-gradable adjective).

From the above summary we can surmise that there are various undiscovered constraints to word co-occurrence other than the semantic prosody theory and the maximizer–booster distinction. It is important then for the learner to observe collocations at play in real texts and internalize any collocational rules, which seem even more difficult to verbalize than grammatical rules.

4. EXTRACTING COLLOCATIONS

This section introduces the method we used for extracting collocations from corpora. The reason we want to extract collocations from English corpora in this context is to use them as a standard against which to check the learner's input word groups.

A number of methods exist for extracting collocations from a corpus, ranging from very simple to very complicated measures. Barnbrook (1996)

reviews the three most common methods for extracting collocations: *z* score, *t* score, and mutual information. More elaborate methods of extracting collocations from a corpus have also been proposed, usually using more strict criteria; for example, the Xtract of Smadja (1993) and the 'cost criteria measure' of Kita and Ogata (1997).

Because the current aim is to use the Collocation Dictionary to process the second language learner's input (explained later in Section 7), collocation extraction methods which focus on collocational strength are preferred in order to cover entries ranging from very strong collocations to loosely collocated items. The idea is to mark each collocation with a 'collocation strength' score in its dictionary entry. When the learner uses a weaker collocation, the system, while accepting this usage, will nevertheless suggest stronger collocations with similar meaning in order to build on the learner's collocational knowledge and add to her fluency. In this context, the *z* score method described in Barnbrook (1996) seems appropriate.

In our task the 100 million word British National Corpus (BNC)[2] is used as the reference corpus from which the standard English collocations are extracted. In order to extract V–N collocations a three-word 'half-span' (i.e., three words to the right or left of the node word, depending on whether the verb or the noun is the current node word) is used. This is to allow maximally an article and an adjective between the verb and the noun. The precision measure will be low if we allow a larger half-span, because then more structures other than the verb–object type would be included which are not relevant to this particular setting.

To demonstrate, if we want to extract collocates for *confidence* from BNC, using a concordancer we get a frequency table of verbs within three words to the left of *confidence* (see Table 1).

There are altogether 6,790 tokens of *confidence* in BNC (i.e., 6,790 concordance lines); with a half-span of three words this means we are dealing with a $3 \times 6{,}790 = 20{,}370$ word extract in which to investigate the significance of a potential verb collocate of *confidence*. Suppose we want to decide whether *destroy* collocates with *confidence*. We calculate the *z* score for *destroy* as follows:

$$z = \frac{O - E}{\sigma}$$

where:

2. See http://info.ox.ac.uk/bnc/ for more information on BNC.

O = the observed frequency for a potential collocate appearing within the designated span of the concordance extract. According to Table 1 there are 10 tokens of *destroy* in this extract, so $O_{destroy} = 10$.

E = the expected frequency for a candidate in the extract. There are 1,978 tokens of *destroy* in BNC, so the expected frequency of *destroy* in the 20,370-word extract is:

$$E_{destroy} = \frac{1978}{100000000} \times 20370 = 0.403$$

σ = the standard deviation of candidate in the corpus,

$$\sigma = \sqrt{\mathrm{N} \times p(1-p)}$$

where:

N = number of tokens in the extract

p = probability of candidate in the whole corpus, i.e.,

$$P_{destroy} = \frac{1978}{100000000}$$

Therefore:

$$\sigma_{destroy} = \sqrt{20370\left(\frac{1978}{100000000}\left(1 - \frac{1978}{100000000}\right)\right)} = 0.635$$

so:

$$z_{destroy} = \frac{10 - 0.403}{0.635} = 15.11$$

According to Barnbrook, a z score of over 3 is worth considering a collocate for the node word. For the system being proposed, collocations crossing the threshold of 3 will be recorded along with its z score, signifying the strength of the bond. When requested, collocations with similar meaning but having higher collocational scores can be retrieved and presented to the learner when a lower score collocation is found.

5. TESTING THE REALITY

In order to justify the use of corpus processing results as the teaching norm, a questionnaire was designed to test native speakers' intuition regarding some verb–noun collocations. The questionnaire contains fifteen questions, each of which presents four candidate verbs followed by a noun in the context of a sentence, as follows:

> He said he had already _____ a complaint.
> ()lodged ()made ()reported ()submitted

The subject is asked to rank the suitability of the verbs in such a context. The *z*-score for each verb–noun pair appearing in the questions was calculated from BNC beforehand. The idea is to compare the *z*-scores for these collocations with the ranking given by the native speakers and see if they largely coincide. In this experiment, 119 native speakers did the questionnaire. Table 2 shows the result of this part of the survey.

In the BNC column, the ranking of verbs is based on the *z*-scores calculated from the previously mentioned method. In the native speakers column, the ranking is determined by the collective scores given by the 119 subjects to each verb. The percentage of each verb reflects the scores it received in relation to the set of four verbs in each question.

It can be seen from Table 2 that the two sets of ranking reveal very good similarity. Exactly the same ranking is obtained in seven out of fifteen questions for the two groups (1, 4, 6, 7, 11, 13, 15). For three further questions (3, 9, 14), the same first and second choices are maintained. For the remaining five questions, three (2, 5, 12) are given the same first choice by the two groups, and, for the other two (8, 10), the first two choices are correctly separated out from the lower ranked ones (third and fourth) even though their order is reversed. Altogether, this seems to indicate that the collocation strength calculated with the *z*-score method is strongly supported by the native speakers' intuition.

Besides letting native speakers do the ranking, we also collected questionnaire answers from 40 Chinese speakers and 31 other non-native speakers (mostly European language speakers), all advanced learners currently studying in a UK university. Collective ranking scores similar to that done for native speakers in Table 2 have been calculated for these two groups. To simplify the discussion, the only fact to be mentioned here is a comparison between the

Table 2. Comparison Between Corpus Processing and Native Speakers' Intuition.

Noun and candidate verbs	*z*-scores from BNC and ranking	Native speakers' collective ranking
1. complaint		
()lodged()made()reported()submitted	1(113.78) 2(21.86) 4(0.45) 3(5.5)	1(37%) 2(36%) 4(7%) 3(20%)
2. activity		
()do()execute()perform()practise	4(–9) 2(–0.4) 1(3.86) 3(–0.6)	2(26%) 4(12%) 1(44%) 3(18%)
3. issue		
()address()cover()examine()treat	1(75) 4(–0.4) 2(6.29) 3(0.2)	1(39%) 3(18%) 2(33%) 4(9%)
4. confidence		
()acquire()gain()get()obtain	2(4.1) 1(63.6) 4(–0.6) 3(1.52)	2(29%) 1(45%) 4(12%) 3(13%)
5. information		
()convey()deliver()transfer()transmit	1(53.3) 3(9.17) 4(–2.6) 2(29.81)	1(44%) 2(27%) 3(15%) 4(14%)
6. relationship		
()build()construct()establish()make	2(11.71) 3(0.65) 1(38.26) 4(–1.2)	2(36%) 3(14%) 1(43%) 4(7%)
7. response ()arouse()elicit()extract()stimulate	3(2.96) 1(92.72) 4(0.97) 2(8.6)	3(18%) 1(41%) 4(17%) 2(25%)
8. trouble		
()cause()generate()make()stir-up	2(64.46) 4(–0.7) 3(6.48) 1(128.13)	1(38%) 4(6%) 3(27%) 2(29%)
9. agreement		
()achieve()get()make()reach	2(10.02) 4(1.32) 3(1.79) 1(124.2)	2(27%) 3(11%) 4(8%) 1(54%)
10. research		
()do()conduct()perform()undertake	4(–12) 2(49.65) 3(2.88) 1(58.45)	3(17%) 1(39%) 4(16%) 2(28%)
11. experience		
()acquire()get()gain()obtain	2(7.04) 4(–2.5) 1(63.32) 3(1.9)	2(25%) 4(16%) 1(40%) 3(19%)
12. example		
()cite()make()quote()take	1(60.52) 4(–1.7) 3(7.85) 2(35.84)	1(40%) 4(11%) 2(30%) 3(19%)
13. service		
()conduct()execute()perform()provide	3(6.77) 4(–0.6) 2(16.71) 1(74.83)	3(11%) 4(6%) 2(31%) 1(53%)
14. knowledge		
()absorb()acquire()learn()obtain	3(–0.6) 1(74.52) 4(–0.8) 2(5.83)	4(7%) 1(46%) 3(8%) 2(38%)
15. success		
()assure()ensure()guarantee()warrant	3(2.65) 1(32.84) 2(31.87) 4(–0.64)	3(11%) 1(42%) 2(39%) 4(7%)

three groups regarding the correctness of their first choice in relation to the BNC. As can be seen from Table 2, the native speaker group has 13 correct first choices when compared with BNC. According to our statistics (not shown here) the European non-native speaker group has 10 correct choices, whereas the Chinese speaker group has only 7.[3]

Overall, the experiments suggested that the Chinese learners of English need extra help in mastering English collocations, and that using the *z*-score method to extract English collocations from corpora provides a good basis for doing this.

3. A different marking scheme was also used, which calculated each subject's scores to be applied in ANOVA tests, which found the Chinese speakers' English collocation knowledge to be significantly inferior to that of European language speakers, which was in turn significantly inferior to that of native speakers of English. A paper addressing this experiment in full is in preparation. For some details see: http://www.dai.ed.ac.uk/daidb/people/homes/shei/survey.html.

6. LEARNER CORPUS

Two kinds of corpora are important in the language teaching context: reference corpora and learner corpora. A reference corpus is necessary for making generalizations or drawing examples of native speakers' language behaviours. A learner corpus is needed for diagnostic purposes, among others.

A learner corpus of English texts written by post-intermediate Chinese learners of English has been collected and is constantly being expanded. The initial work to be done with the corpus is to develop an Error Library based on it which consists of unacceptable collocations by native speakers.

As yet no algorithm that the authors are aware of deals with automatically extracting incorrect collocations from learner corpora. The problem is different from extracting collocations from native speakers' corpora because, while we are interested in finding the more frequent erroneous collocations, we are equally interested in spotting any single appearance of anomalous collocation which the learner may produce.

The first stage of collocational error analysis deals with the verb–noun type collocation, which as mentioned before is found by Gitsaki to be the most difficult for learners. A 'V–N extract comparison' method is used. First, the entire learner corpus is tagged using the LT part of speech tagger, a program obtained from the Language Technology group of HCRC.[4] Secondly, a trivial computer program is written which can extract all words tagged as *verb* from the corpus, followed by a noun within a certain span. Each V–N pair is then compared with the V–N pairs extracted from BNC which have the same leading verb. A V–N pair from the learner corpus having no counterparts in the BNC will be marked as suspicious. Manual inspection is then conducted to verify that this is indeed an illegal combination. If so, this is entered into the Error Library. Some incompatible verb–object combinations exposed by this method include *make action, did effort, cease confidence,* and *perform help*, etc.

7. SYSTEM DESIGN

In this section the main function of the Corpus-Based Collocation Tutoring System is described and its main architecture is presented (Fig. 1). As the focus is on the targeting of suspicious collocations in the user's input, other aspects of the system (e.g., the concordancer and the tutoring and student modelling modules) are ignored in this discussion.

4. See http://www.ltg.ed.ac.uk/software/ for details.

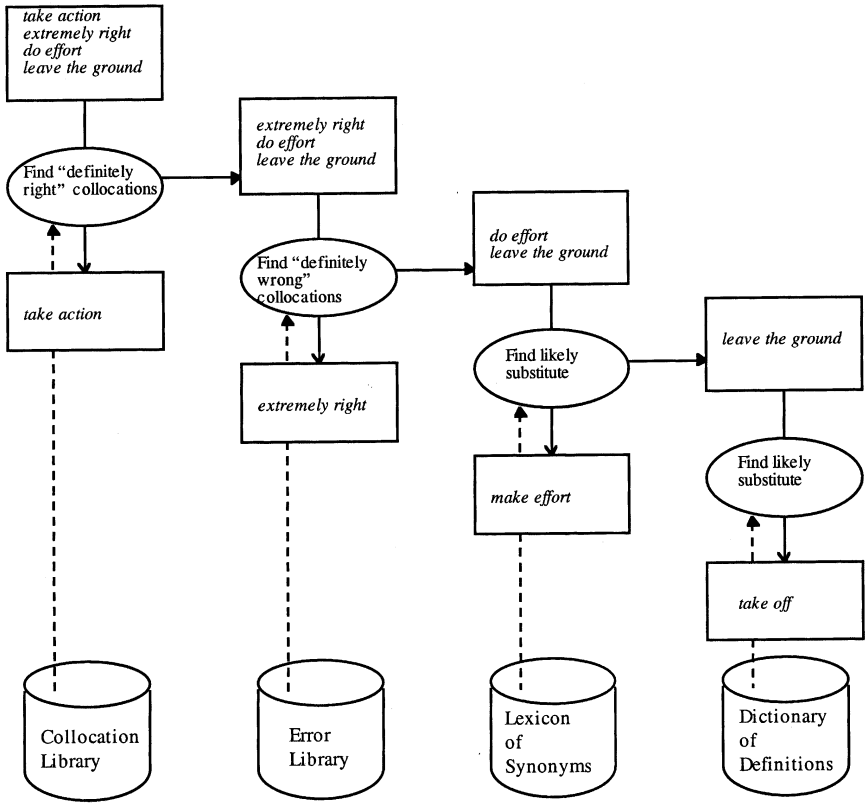

Figure 1. The Corpus-Based Collocation Tutor.

7.1. System function

The system takes a stretch of English text input by the student, detects any potential collocational errors, and offers example concordance lines from the reference corpus which contain the correct usage of the collocations in question. For example, when a student writes *make action* and receives a warning message from the system, the student will be prompted to check the concordance lines containing *take* and *action* in order to generalize the correct usage of the relevant collocation.

A writer's collocational aid should aim to do more than just point out the errors, however. Many post-intermediate ESL learners write grammatically correct English but their writing often lacks a native-like flavour or is not

versatile in diction. This could be partially due to inadequate collocational knowledge. It is useful to point out to these writers which word sequence (like *different in all respects)* in their writing could be replaced by which collocation (say *fundamentally different*), examples of usage being provided by concordance lines drawn from the database. This is more difficult to do but is very valuable in enhancing ESL learners' collocational knowledge, since learners frequently use an 'avoidance' strategy in production, using only words and structures they are confident with while avoiding unfamiliar ones (see Ellis, 1994, p.304, for a brief discussion on avoidance).

7.2. Description of modules

7.2.1. Collocation Library
This is a collection of English collocations covering as comprehensively as possible the structures currently being investigated. For example, if our focus is on V–N collocations, then the Collocation Library should include a long list of items such as *achieve goal, hire car, make friends, take action*, etc.

7.2.2. Error Library
This is a collection of collocational errors previously uncovered by studying learner corpora. This keeps expanding as new errors are discovered through the use of the system by the learner. For each collocational error, the correct collocation form(s) will also be stored for retrieval as evidence of correct usage.

7.2.3. Lexicon of synonyms
This is like a thesaurus, except that it contains as entries only the words being incorporated in the Collocation Library. So, if we have *achieve goal* in the Collocation Library, then we will have a group of words associated with each of these two words listed under their separate heading, like so:

achieve: get, gain, accomplish
goal: aim, objective, purpose

7.2.4. Definition Dictionary
The content of the Definition Dictionary is decided entirely by its functionality. Each entry is headed by a collocation in the Collocation Library, followed by a number of words. The words are of two categories: they are either 'relatives' or 'neighbours' to the collocation elements. Relatives include synonyms of each

collocation element and, more broadly, words of relevant concept. Neighbours include, in the case of V–N collocation, the likely subject noun frequently associated with the collocation. An example entry in the Definition Dictionary is:

isolate issue
relatives: divide, ignore, problem, independent, . . .
neighbours: we, I, . . .

The inclusion of synonyms is straightforward, whereas that of 'relevant concept' words and neighbours words is more problematic. At this stage, it was conceived that the former should be derived from a 'corpus of paraphrases' gathered from the learners themselves. The idea is first to teach the collocation to the students, and then to ask them to use words they know to define or explain it. For example, for the collocation *isolate issue*, we may get definitions from learners like: *to ignore other problems, to deal with problems at different times*, etc. Thus when we have a large number of learners doing the task, we can collect a corpus of paraphrases for different collocations. We can then extract words of relevant concept from it—e.g., *ignore, deal with* and *different* from the above example—to be included in our Definition Dictionary. When the system processes a learner's input writing, the opposite procedure is applied: a better collocation is provided in place of less mature learner phrases.

The inclusion of neighbours in a collocation, on the other hand, comes from a native speaker's corpus like BNC. A noun or pronoun whose *z*-score for collocating with the collocation in question passes a certain threshold will be recognized as a neighbour to this collocation in the Definition Dictionary.

7.3. Processing

7.3.1. First stage

The system tags and chunks the user's input and picks out the elements of the structures in question, for example the V and N of a VP, or an Adv followed by an Adj.

7.3.2. Second stage

These structures are first compared with the Collocation Library to see if they match any legitimate collocation. If so, either they are passed without further processing or, if the learner requires, they go through a 'synonymous collocation finding' process, where collocations with similar meaning but with a higher *z*-score are produced.

7.3.3. Third stage
The groups of elements are compared with the corresponding categories in the Error Library. If an exact match is found, it is singled out. Correct collocation forms are provided with example concordance lines illustrating their usage.

7.3.4. Fourth stage
The structures which are neither definitely correct nor definitely wrong are considered at this stage. This is an interactive process between the user and the system. Here the dictionary of synonyms comes into play. For a given suspicious collocational pair, the system searches each element in turn for its synonyms, comparing each of these with the other element of the pair to see if they form a legitimate collocation. If so, the correct form is suggested, supported by concordance lines to allow the user to compare and decide whether this is the form they want. For example, for the pair *eat medicine*, which is mistaken but is new to the Error Library, we hope to find *take* as a synonym of *eat*, which will match with *medicine* and find itself as a legitimate entry in the Collocation Library. The system will then suggest *take medicine* to the user.

7.3.5. Fifth stage
The final stage of processing looks for structures in the learner's input which could be replaced by more native-like or rhetorically useful collocations through the 'definition' method. Two possible levels of functionality are involved here: first, tidying up unfavourable phrases, like changing *The plane left the ground* to *The plane took off;* second, providing collocational alternatives for otherwise legitimate constructions: for example, suggesting *Help is sorely needed* in place of *Help is needed very badly*.

To illustrate how the mechanism should work, consider a sentence in our learner corpus:

I suggest that we make two issues independently

At this stage, not only the words corresponding to the collocation of concern are extracted, but the neighbouring (say, two) words are also taken into consideration. So, from the above example, *that we make two issues independently* is extracted. In the case of V–N collocation, the noun is usually central to the structure, so it should be considered first. The system looks for entries in the Definition Dictionary which contain this word. When it comes to the definition of *isolate issue*, several matches are found in the categories of

neighbours and relatives: *we*, *make*, and *independently*. Ideally each item in the dictionary entry will carry a weight derived from learner experiment data and BNC as discussed in the previous section. The matched collocations can then be presented to the learner in appropriate order according to the scores they obtain.

Note a meaningful side-effect of stage four processing is, when the user actually accepts the correction of a problematic collocation, the system can 'learn' a new collocational error, which can be added to the Error Library, which thus keeps expanding. The processing cost of the system becomes lower as more errors are spotted at an early stage of processing.

8. CONCLUSION

In this article the importance of teaching English collocations to second language learners has been discussed. More importantly a method has been proposed for using native speakers' corpora and learner corpora to direct the design of an language tutoring system, starting from the extraction of norms from native speakers' corpora and the computer-aided error analysis of learner corpora. The architecture of a corpus-based intelligent tutoring system currently being developed has also been described, which hopefully will guide learners toward better writing in English by pointing out their potential errors in collocation and allowing learners to observe the use of collocations in authentic language.

REFERENCES

Aston, G. (1995) 'Corpora in language pedagogy: Matching theory and practice', in G. Cook & B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson.* Oxford: Oxford University Press.

Altenberg, B. (1991) 'Amplifier collocations in spoken English', in S. Johansson & A. Stenström (eds) *English Computer Corpora: Selected Papers and Research Guide.* Berlin: Mouton de Gruyter.

Barnbrook, G. (1996) *Language and Computers: A Practical Introduction to the Computer Analysis of Language.* Edinburgh: Edinburgh University Press.

Cruse, D.A. (1986) *Lexical Semantics*. Cambridge: Cambridge University Press.

Ellis, R. (1994) *The Study of Second Language Acquisition.* Oxford: Oxford University Press.

Garside, R., Leech, G. & McEnery, T. (eds) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Harlow, Essex: Addison-Wesley.

Gitsaki, C. (1996) 'The development of ESL collocational knowledge', Ph.D. thesis. Brisbane, Australia: Center for Language Teaching and Research, The University of Queensland.

Johansson, S. (1993) '"Sweetly oblivious": Some aspects of adverb–adjective combinations in present-day English', in M. Hoey (ed.) *Data, Description, Discourse: Papers on the English Language in Honour of John McH Sinclair.* London: Harper Collins.

Kita, K. & Ogata, H. (1997) 'Collocations in language learning: Corpus-based automatic compilation of collocations and bilingual collocation concordancer', *CALL* 10 (3): 229–38.

Kjellmer, G. (1991) 'A mint of phrases', in K. Aijmer & B. Altenberg (eds) *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* Harlow, Essex: Longman.

Smadja, F. (1993) 'Retrieving collocations from text: Xtract', in S. Armstrong (ed.) *Using Large Corpora*. Cambridge, MA: MIT.

Stubbs, M. (1995) 'Corpus evidence for norms of lexical collocation', in G. Cook & B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics: Studies in Honour of H.G. Widdowson.* Oxford: Oxford University Press.

Zhang, X. (1993) 'English collocations and their effect on the writing of native and non-native college freshmen', Ph.D. thesis, Indiana University of Pennsylvania.