Collocation and Text

Hiroaki Otani
University of Tokyo

1. Research goal

The aim of this research is to see whether certain collocations are activated for any given topic and also to see the differences and similarities within each group (i.e. native speakers and learners) as well as between the two groups. By doing so, we will be able to observe how and to what extent native speakers' utterances actually consist of prefabricated patterns as Pawley and Syder (1983) suggested.

2. The Corpora

To achieve the research goal described above, mini-corpora consisting of texts of native speakers of English and Japanese learners of English were created in the following way:

[Instructions for the essay]
Write an essay of about 100 words on the following topic:
Do you prefer to shop at large department stores or small individual shops? Why?

[native speakers' corpus]
50 texts 5541 words
nationality of subjects: British (41) American(8) Australian (1) (avg. age 35.6)
avg. number of words per text. 110.8 words
Text written as an answer to a questionnaire (with almost unlimited free time)

[learners' corpus]
61 texts 5552 words
high school student (in the second year) who learned English as foreign language for five years (avg. age:17) avg. number of words per text: 91.9 words
Texts were written as an answer for a test (under certain pressure of time and without consulting any reference books as dictionaries.

---

3. Analytical approach

In this section, I would like to explain the analytic approach I have adopted for this research. Perhaps the most usual way to investigate collocational patterns is to compare KWIC concordance lines of several chosen keywords between two groups. For instance, we may study concordance lines of the keyword *price* within each group as well as between two groups, since this word is highly likely to occur under our present topic. However, this method has a serious limitation in that any such keyword can only be chosen in a highly arbitrary and accidental way and the scope of results available is thus very limited. Actually, nobody can predict precisely what kind of words or collocations will actually be called forth in a particular subtopic. All we can do is to make random guesses: one can only say that such words as *cheap, expensive, price,* or *bargain* might be included in a text discussing shopping but this can by no means be an accurate list.

To avoid this problem, I have adopted the following approach: taking advantage of relatively small corpora, I read all the texts carefully and gathered all the patterns, in a heuristic way, which went into the expressions of several subtopics. Of course, this procedure, too, can be arbitrary in the categorization of subtopics, but our main goal here is not to make precise categories but to discover and gather typical collocational patterns appearing in texts. Once these patterns have been successfully collected, we can apply various analytic strategies to these results as in normal research of this kind. We will also discuss these results against the data from a more general corpus, that is, British National Corpus (BNC hereafter).

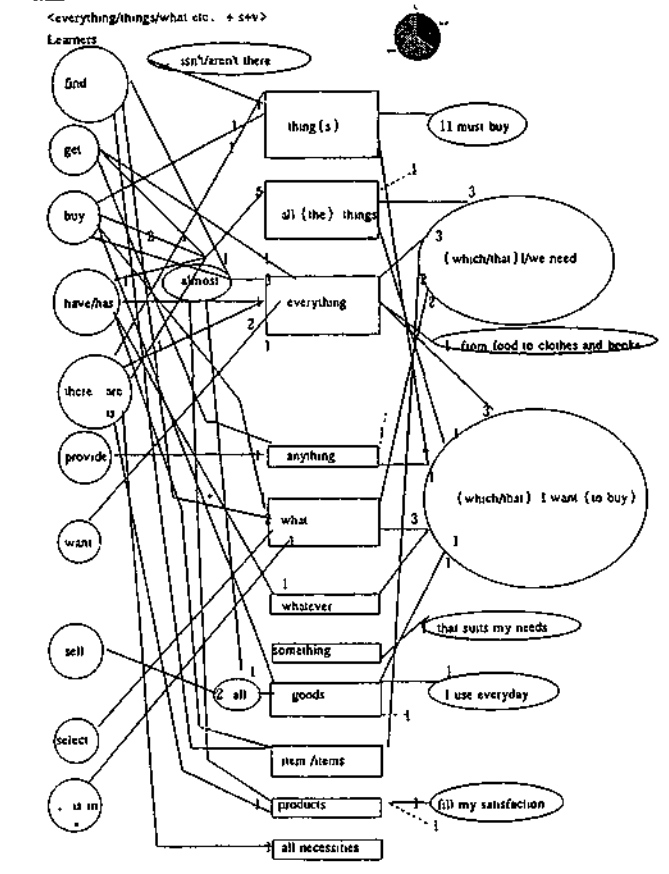
The subtopics we are dealing with concern (A1) quantity and variation of products available at a store, and somewhat related to this, (A2) the advantage of a department store in that one can do all one's shopping at one place, and finally (B) description of the relationship between customers and either type of store. Typical and recurrent patterns in each subtopic are shown in charts as in the figures in the following section. Each chart is complete with a pie chart showing a proportion of subjects using each pattern so that one can see to what extent each pattern is shared among different texts. Also, the number in a chart indicates the number of subjects using each word.

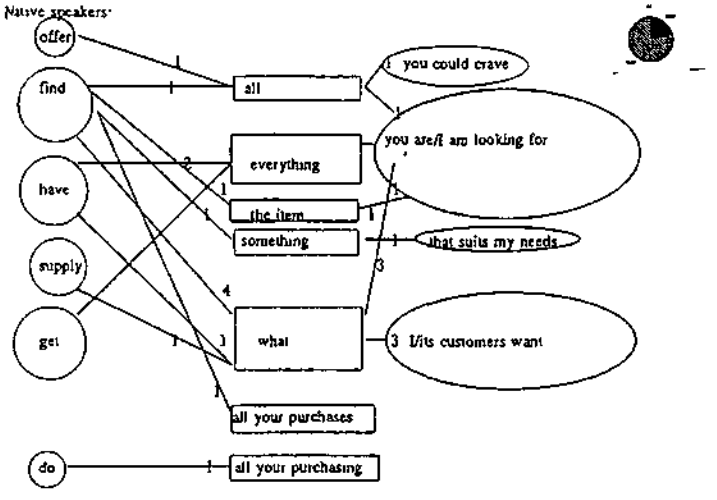
4 Results

4.1 (A1) Expression for 'range' of products

In order to talk about the great variety of products on sale at department stores (which is normally the advantage of shopping in such places), one of the most transparent options may be to say something like *there is everything I may want at department stores* or *I may find what I want at department stores.* As we look at Figure 4.1a below, we find that although such patterns are observed both among learners and native speakers, they are especially frequent among the former. Besides, there seems to be a slight difference in preference between the two groups using this pattern: while learners prefer 's+ want' in what-clause or relative clause following the head noun *everything*, native speakers show a preference for *be looking for,* though *want* is found in both groups. It is interesting to note that even in such a minimally idiomatic and transparent frame, certain 'local' difference in preference does seem to appear between learners and native speakers. At the opposite pole of the expressions will be a highly lexicalized, concise pattern like *find all your purchases* and *do all your purchasing* which are only found in native speakers' texts. However, this type of 'short-cut,' option seems marginal, even among native speakers.

Figure 4.1a

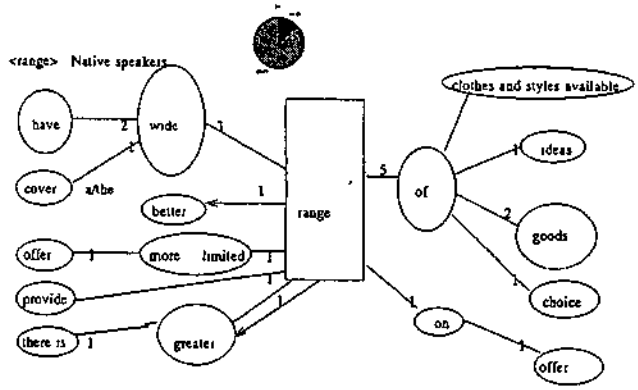




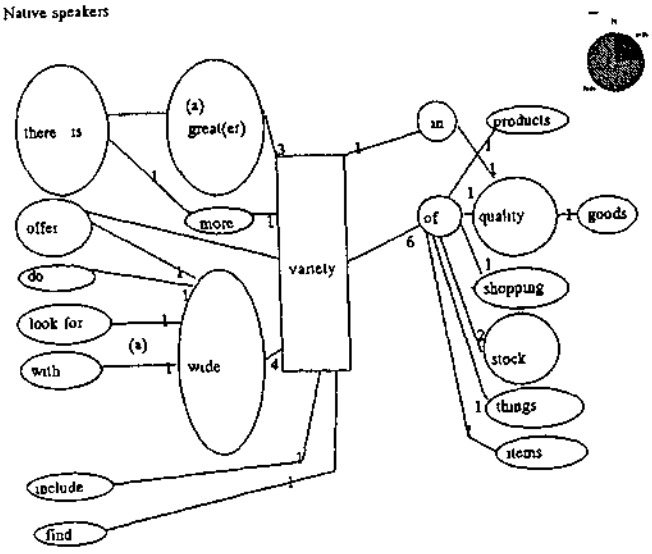
The most frequent pattern, which is found almost exclusively among learners (only one instance by a native speaker), is *there are many things at department stores* or *department stores have many things* type sentences. Indeed, nearly half the learners use this type. However, this seems hardly surprising, as it will probably be the easiest and the most transparent option to refer to this content. Here, paradoxically enough, the lack of knowledge of fixed patterns results in yet another highly fixed pattern.

Patterns containing such keywords as *variety, range, selection, choice* are salient among native speakers (Figure 4.1b). (There is no instance of these words except *choice* among learners, whose counterpart is *kind*.) If we look carefully at patterns in which these four keywords appear, we may notice that there are striking similarities among them. Apparently native speakers here seem to choose one element from each group (somewhat in a paradigmatic way) and combine it according to a shared syntagmatic framework, even if these elements are not interchangeable to a full extent (Table 4.1c). Like a person choosing one item from several different categories in a cafeteria (for example, a salad, a main dish, and a drink), choice is apparently made within a more or less closed set of items and based on a shared syntagmatic container. In other words, all the words within this group (*wide, great, large; range, variety; of*) fit neatly into a certain shared syntagmatic and semantic schema. When we consult a large general corpus (BNC) (Table 4.1d), on the other hand, we find these patterns are again highly typical even in the overall use of each keyword. However, there are more varieties in the nouns following a preposition *of*; they are not necessarily such nouns as *products* or *items* but a 'wide range of' nouns' as it were, belonging to schemata other than of shopping, for example *a wide rage of activities/businesses/excuses etc.*, along with such words as *things, products*, and *items* which are strongly suggestive of a similar schema to ours. Still, it could be said that all these nouns seem to share a yet larger schema in which an utterer tries to let his or her audience admire something for its variety (perhaps sometimes in an ironical context as in the case of *excuses*).

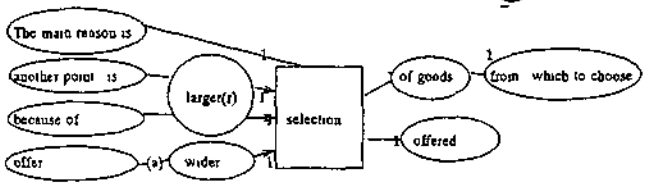
Figure 4.1b



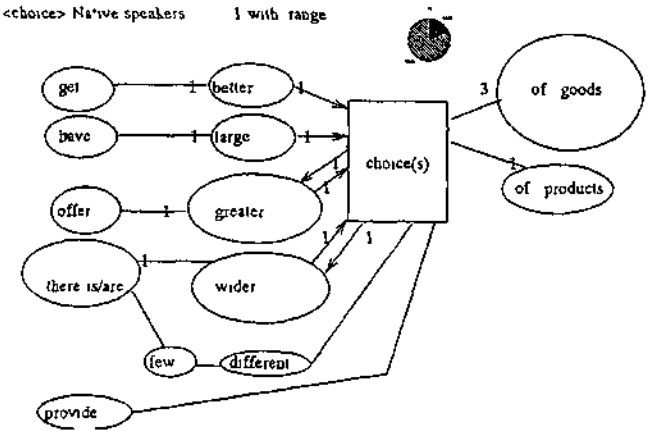
Native speakers



<selection> Native speakers



<choice> Native speakers 1 with range



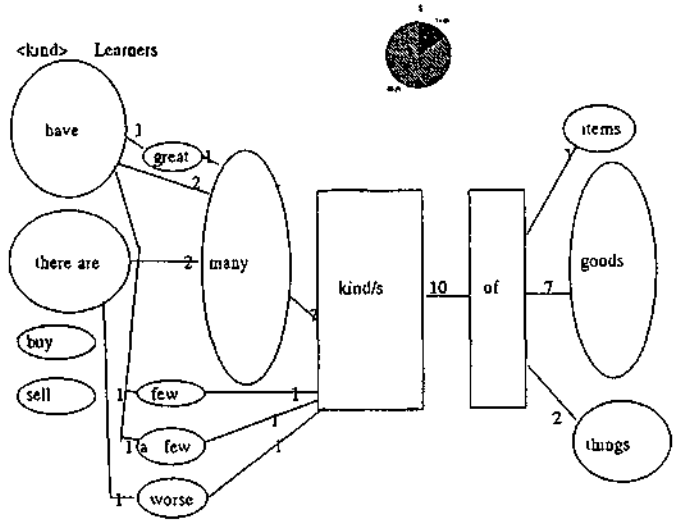


Table 4.1c

Syntagmatic framework [a+Adj+N1+of+N2]
Likely fillers for Adj position:wide(r), great(er), better, large(r),
Likely fillers for N1 position:range, variety, selection, choice
Likely fillers for N2 position: goods, things, products, items, choice

---

Table 4.1d

How typical are the expressions used in our native speakers' texts (on a specific topic) in a large, general corpus (BNC)? (Numbers show number of instances in 100 random occurrences (one per text)in the corpus) Those without number occurs only once.)

[range]	V+Adj+ range of+N(+available)			
there is(2)	*wide(5)*	*range of (64)*	*goods(2)*	*available(3)*
offer (2)			*service(s)(2)*	
provide			*things(3)*	
supply				
have				
[variety]	V+Adj+variety of+N			
	wide(7)			
offer	*great(2)*	*variety of(63)*	*colours(3)*	
provide	*large(2)*			
[selection]	V+Adj+selection of +N			
there is	*good(3)*		*clothes*	
have	*large(3)*		*products and service*	
	wide(2)	*selection of((49))*	*shops and restaurants*	
	small			
	varied			
	broader			
[choice]	V+Adj+choice of			
offer(4)	*much(3)*		*(No noun relevant to*	
there is	*more(2)*	*choice of ((35))*	*our topic)*	
get	*free*			
have				

If we integrate these patterns of four keywords above:

have	*large, small, great*	*range*	*products*
there is	*wide*	*variety of*	
provide		*selection*	*service*
offer		*choice*	

4.2 (A2)*Under one roof*

One of the biggest advantages department stores offer may not only be that they provide a wide range of items but also that they are usually housed in one building. It is found that in making this point, a fixed idiom, *under one roof* is often called forth, exclusively across native speakers' texts (Figure 4.2a). In fact, this expression is used by 9 different subjects, which amount to nearly 20% of the native speaker subjects. Furthermore, the structures which co-occur with it are found to be more or less similar. Probably we will be able to say that the highly invariable, fixed pattern *under one roof* seems to be again a part of a yet larger loosely fixed (but fixed nevertheless) pattern. Even more important, it is likely that *under one roof* is already closely related to, or strongly suggestive of, a certain very specific schema, that is, 'shopping at large stores.' Looking at how the expression *under one roof* actually occurs in a large corpus (Table 4.2b), we can find that many samples seem to be used in strikingly similar contexts to ours. In fact, similarities are such that some sentences in BNC seem perfectly transferable to our native speakers' texts. To put it briefly, the phrase *under one roof* and the loosely fixed pattern surrounding it are sometimes directly linked with, or rather a part of, a specific schema of shopping. (Otherwise, much less frequently, *under one roof* is used with the word *family*) In fact, since the phrase *under one roof* is very often repeated in commercial discourse (mainly in advertisements), it is quite natural that people should be automatically conditioned to associate this particular expression with a particular topic.

Fig 4.2a

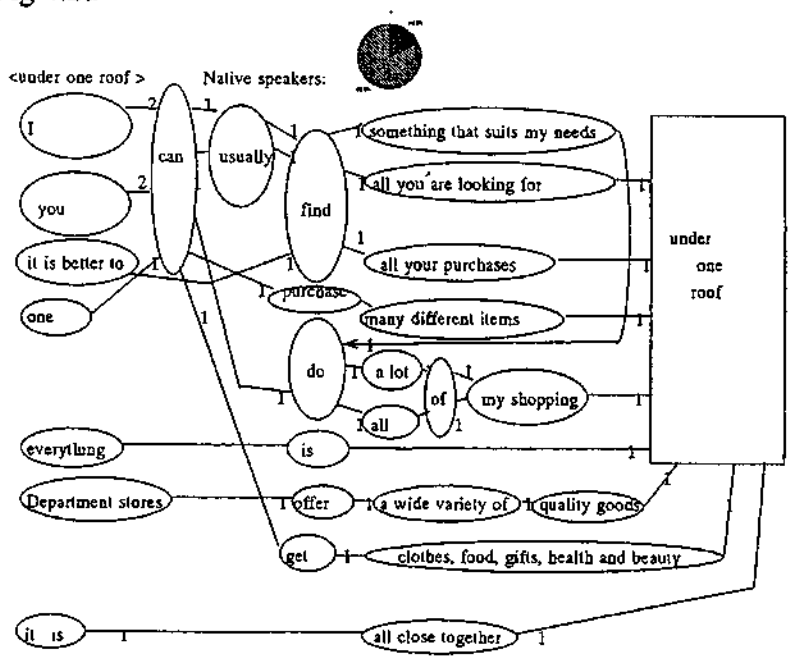


Table 4.2b

under one roof in BNC
*...everything is **under one roof***
*...where everything could be purchased **under one roof***
***Under one roof** you could buy food, clothes and electrical goods and park just outside.*
*...you can shop throughout the store, visit every floor, buy all your gifts **under one roof***
*...computer superstores allow the undecided or the uninitiated to see **under one roof** what is on offer and to seek on-the-spot advice*
*...which offer free car parking and hundreds of shops **under one roof***
*...which include many retail outlets **under one roof***
*Department stores have a variety of different departments **under one roof.***
*These are stores that have five or more departments **under one roof***
*...The rest of the afternoon we spent shopping and taking in the experience of having so many shops and stalls **under one roof.***
cf. under one roof in our native speakers' texts
*...you can usually find all you are looking for **under one roof.** (n6)*
*...everything is **under one roof** and prices are more competitive. (n11)*
*...it is better to find all your purchases **under one roof.** (n12)*
*The fact that one can purchase many different items **under one roof** is... (n18)*
*...I can do all of my shopping **under one roof** (n20)*
*I can usually find something that suits my needs and also do a lot of my shopping **under one roof.** (n1)*
*Department stores offer a wide variety of quality goods **under one roof** (n32)*
*You can get clothes, food, gifts, health and beauty all **under one roof.** (n44)*
*...in a department store it is all close together **under one roof.** (n50)*

4.3 (B)Expressions describing the relationship between customers and either type of a shop

Next, we will look at what kinds of expressions are used in order to describe a more abstract idea: the relationship between small shops or department stores and their customers. First, learners mainly use such words as *friendly, friend* (as part of an idiom, *make friends with*), *kind* (adjective), and *familiar*. Among these words, only *friendly* is shared by native speakers. As regards native speakers' texts, on the other hand, there are several typical, recurrent patterns to be observed. We can find such words as *individual, faceless, personal, face to face*, and *anonymity* in use. At a glance, however, these seem to have little in common with each other unlike in the case of *range* or *variety*. However, two adjectives *personal* and *individual* do share many semantic features even at dictionary definition level in that both words highlight an individual person. The use of these words, therefore, suggests that those who use them are somewhat inclined to define the relationship between a customer and a shop in terms of an 'individual person.' From this perspective, such words as *faceless, anonymity* (or the adjective *anonymous*) which are found in our native speakers' texts can also be seen as the opposite of words such as *individual* and *personal*. In fact, the original meaning of both *faceless*, and *anonymity* is 'absence of a person.' That is, both *face* in *faceless* and *onymo* (Greek meaning 'name') in *anonymous* are metonymies representing a 'person,' and the suffix *-less* and prefix *a--* both indicate absence. Although etymological criteria can be misleading, it would be undeniable that these two words do have such semantics even synchronically, regardless of their origins. Interestingly, one native speaker subject happens to write in such a way: *the lack of personal service* (in department stores), which can directly 'translate' as

anonymity or *facelessness*.

In this light, such expressions as *face to face* and *one to one* used by our native speakers can also be seen as an extension of *personal* and *individual*, in that both are reducible to 'person to person' and then finally to a single lexicalized adjective *personal*, or *individual*. Furthermore, the former expression *face to face* is obviously an antonym of *faceless*. What is important is that, in our corpus, *faceless* and *face to face* occur not in the same text but across different texts.

These results do highlight an important feature in the selection of words by native speakers referring to this particular subtopic: the relationship between a shop and its customers is likely to be discussed in terms of a 'person.' Perhaps many native speakers are not quite aware of this, since any arbitrary articulation tends to seem inevitable to its utterers once it is established *within* the system of language. However, in our contrastive study, we can see that this conceptual and schematic reference to 'person' is unique to native speakers and it is quite recurrent and typical only *within* this group. Although I am empirically quite sure that almost all high school students who participated in my study know the adjectives *personal* and *individual*, the important fact is that none of them thought of using either of these words in discussing this subtopic.

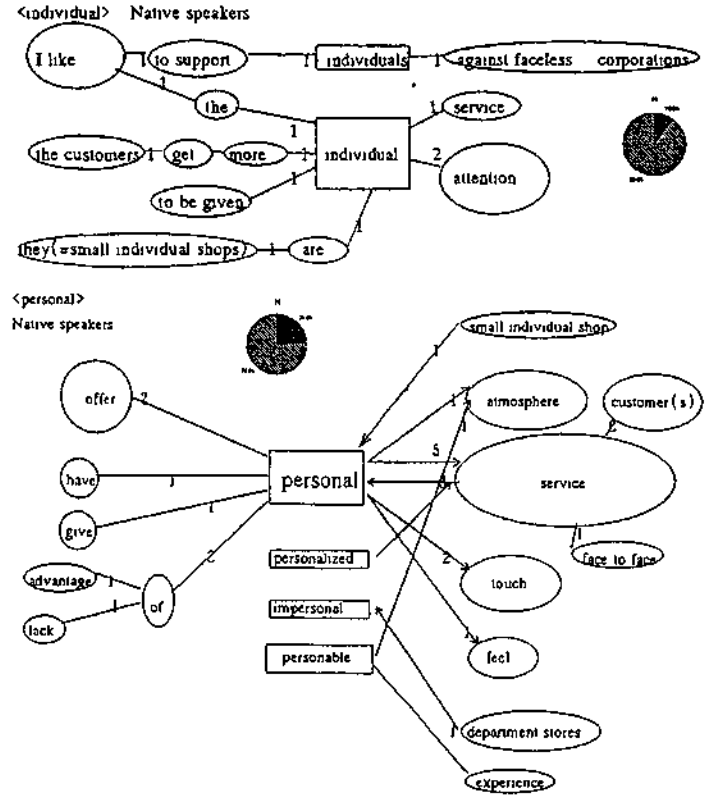
Apparently, in an actual text, several words potentially belonging to this 'person' schema are chosen and put into use. In fact, such words as *personal, individual, anonymous, faceless, face to face,* and *one to one* which appear across our native speakers' texts all seem to fit into a coherent semantic framework. Perhaps compared to the case of *range* or *variety*, this schema is much more conceptual in nature with fewer syntagmatically common features. Learners' preferences for words, by contrast, are generally adjectives referring to human nature, and seems less coherent as a whole.

Native speakers:		Learners:
individual		friendly
personal	anonymous	many people
face to face	faceless	familiar
one to one		kind
friendly		can talk to the clerks

Finally, we will look at the collocational environment of several of these recurrent keywords among native speakers both in our native speakers' corpus and in BNC (Figure 4.3a and Table 4.3b respectively). *Personal,* which is highly likely to be connected with *service* in our native speaker corpus, is seen to have the strongest collocational tie with the word *computer* in BNC and there is one instance of *personal service* out of 100 random samples. (Although there are two instances of *personal social service*, this has a different meaning.) Other seemingly relevant collocations in BNC would be *personal attention*, which reminds us of *individual attention* used by our native speaker subjects. Another interesting example in BNC is *personable, personal, face to face encounter* occurring simultaneously in the same clause, although this happens in a religious context, not in a commercial one. In any event, we can confirm also from this fact that these words do belong to the same conceptual schema not merely in analytic terms but also in actual use. When we look at samples of *anonymity* in our native speakers' corpus, we can see that *prefer/like anonymity of* ...is a recurrent pattern. In a general corpus, on the other hand, we can observe that there are three typical groups (A to C) of recurrent collocational patterns and we notice that *prefer/like anonymity* does belong to group A, the most frequent pattern. (In this sense, it could be argued that the use of the keyword *ano-*

nymity or *anonymous* in a text is not so much an idiosyncratic, individual choice as an 'anonymous' pattern!) Finally, let's look at collocational patterns of a word *faceless* in BNC. Although this word *faceless* occurs only once in our native speakers' corpus, it is of great interest in that it clearly forms a part of a schema mentioned earlier. Here, we have a quite interesting list of people or things regarded as *faceless*. Again, while there are several lines of typical collocates (*bureaucrats, murderer, building* etc.) there is no instance, unfortunately, of *department stores*. Perhaps the closest we get to this is *shopping centres and office blocks*. Another interesting fact observed in BNC is that we can see that *faceless* is often juxtaposed with its (near-) synonyms, apparently to reinforce its meaning. Examples of these synonyms are also shown in Table 4.3b. Looking at the list of such co-occurring synonyms, we can find such words as *nameless* which obviously share the same conceptual schema as *anonymity/anonymous* among others. Such evidence again shows that these words do not simply share common semantic features in analytic terms but they are also *actually* thought of together by a text writer and tend to co-occur in real texts.

Figure 4.3a



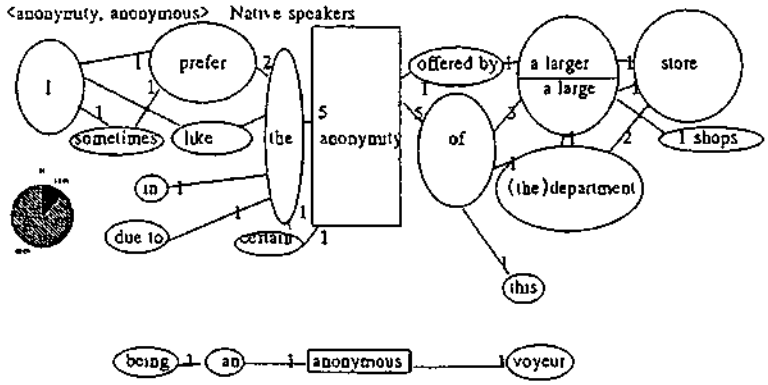


Table 4.3b

[personal] personal+N
service (1/100) (...service industries consist largely of technologically stagnant, small-scale ***personal services...)***
cf. religious festival here is Pilate in this favoured position of having a personable, ***personal*** *face to face encounter with Jesus Christ*
[anonymity] V+anonymity
group A(14/100):*need(2), prefer(2), seek(2), demand, request, desire for, search for, insist on, want*
group B(7/100): *guarantee(2), preserve(3), protect, respect*
group C(7/100): *shelter/hide/hold behind, decay into, sink back into, flee to, escape back to*
[faceless] faceless+N
human(42/100): *bureaucrats(5), bureaucracy(1), Prime Minister, politicians, courts, authorities, financial institutions, academics, experts, murderers(2), killer, agents, director(s)(2), crowd, individuals, spectator, conductor, father, knight, members, planner, steelworkers, New Yorkers, figure(2), man/men(6), woman(2), people(2), someone*
non-human(16/100):*building(2), cement blocks(2), shopping centres and office blocks, architecture, corridor, environment, approach, doll, mannequins, newspaper report, forces, Cheshire cat, pig*
cf. (nearly-) synonymous adjectives juxtaposed with faceless with *and* or a comma(,) in between
nameless(2), heartless(2), hollow, blank, emotionless, empty, featureless
cf. faceless women whose names he probably couldn't even remember

---

5. Summary and conclusion

Certain topics activate certain sets of collocations and a text consists of such patterns (probably prefabs are not an option for the economy of effort or memory but are even a basic unit of language and thought.) However, such collocations tend to occur, to a certain degree at least, not independently but as part of a certain schema. In the case of *range* or *variety*, a paradigmatic choice, as it were, is made within the same syntagmatic framework. In the case of *under one roof*, it has become clear through the analysis of a general corpus that this has a stronger link with a specific schema in which it is likely to occur. Finally as regards the expression referring to the relationship between customers and shops, the keywords used by native speakers are observed to form a coherent conceptual schema across texts. It was also found that when these keywords occur, sometimes other members in the same schema also co-occur in the same texts. Furthermore, the collocational patterns of some of the keywords are found to be quite typical in a general corpus (BNC) as well.

In sum, although certain set of typical collocations do recur for a specific topic, it is not that native speakers use one specific collocation for a specific topic in many cases (the example of *under one roof* is the closest we get to this). In other words, there does not seem to be something like one-to-one correspondence between a specific topic and a specific collocation. If anything, the choice of a word in a collocation (we referred to it as a 'keyword' above) as well as that of other words to co-occur with it is likely to be made from among several possible alternatives in the same schema. Here, although there is a certain degree of freedom in a choice within that schema, the schema itself tends to be very closely linked with a specific topic or a subtopic. What is important here is that such 'schemata' are not artificial ones that are invented in a top-down and purely logical manner for the purpose of analysis or categorization. Instead, these schemata are something that is actually referred to –albeit unconsciously-- when one commits oneself to creating a text, which is therefore only to be discovered through empirical 'field' studies. That such schemata do exist is further confirmed by the fact that those words that are assumed to be components of the same schema sometimes also co-occur in the same text, as well as across different texts. Moreover, such highly recurrent patterns in our specific topic are sometimes also found to be one of the recurrent patterns in a more general, topic-unspecific corpus like BNC. This probably gives us many clues as to how an actual text is made out of 'recycled' or 'deja-vu' prefabs, which will merit further research.

If someone asks a person 'Why do you prefer shopping at department stores?' and he or she replies 'Because I like the anonymity,' it will make perfect sense. I suggest that we should explore what it is to 'make sense' in terms of (potential) prefabs including collocations.

Reference

Pawley, A. and Syder, F. 1983 'Two puzzles for linguistic theory: nativelike selection and nativelike fluency' in J. Richards and R. Schmidt (eds.) *Language and Communication*