Squibs and Discussions

Co-occurrence Patterns among Collocations: A Tool for Corpus-Based Lexical Knowledge Acquisition

Douglas Biber*
Northern Arizona University

1. Introduction

One of the main problems for applied natural language processing is gaps in the lexicon, including missing words and word senses, and inadequate descriptions of word use in context. Traditional lexicography has similar concerns. The availability of large, on-line text corpora provides a straightforward tool for enlarging the stock of words included in a lexicon. The identification of additional word senses and uses is more problematic, however.

Much recent lexicographic work employs concordances generated from text corpora for these purposes. While this approach provides a more solid empirical basis than traditional lexicographic approaches (which depend on the manual collection and sorting of citation index cards), concordances can actually provide too much data. For example, a concordance for the word *certain* produced on an 11.6 million-word subsample of the Longman/Lancaster Corpus generated 3,424 entries; a concordance for the word *right* from the same subcorpus generated 7,619 entries. Simply determining the number of different senses in a database of this size is a daunting task; to accurately group different uses or rank them in order of importance is not really feasible without the use of additional tools.

One such tool is to simply sort concordance lines according to their different collocational patterns. Entries can be sorted according to their collocates on both the left and the right. Many of these collocational pairs show a strong relation to a particular word sense (e.g., contrast *right ear* and *right away*), and thus analysis of collocational relations has become an important tool for lexical knowledge acquisition (see Sinclair 1991; Smadja 1991; Zernik 1991).

In addition, there are statistical tools that can help determine the relative strength of collocational relations. For example, Church and Hanks (1990) describe the use of the mutual information index for this purpose (cf. Calzolari and Bindi 1990). Church et al. (1991) further describe the use of t-scores to assess the extent of the differences between the collocational patterns of nearly synonymous words. These tools are important in that the strongest collocational associations often represent different word senses, and thus 'they provide a powerful set of suggestions to the lexicographer for what needs to be accounted for in choosing a set of semantic tags' (Church and Hanks 1990, p. 28).

However, such tools do not directly characterize word senses or even provide any direct indication of the number of different senses that a word has.[1] Further, these

* Dept. of English, Northern Arizona University, P.O. Box 6032, Flagstaff, AZ 86011-6032; biber@nauvax.ucc.nau.edu.

1 Church and Hanks (1990; Church et al. 1991) thus emphasize the importance of human judgment used in conjunction with these tools.

tools are not designed to assess the relations *among* various collocations, addressing the question of which clusters of collocations reflect similar underlying senses.[2]

The present paper discusses the use of factor analysis (a multivariate statistical technique) as a tool for such research questions. In particular, this technique contributes three types of information not provided by other complementary techniques: 1) an indication of the number of major senses and/or uses associated with a word; 2) an indication of the way that various collocational patterns relate to one another in marking word senses and uses; and 3) a fuller analysis of the senses themselves, based on interpretation of the shared bases underlying the groupings of collocations.

2. Methodology

The use of multivariate statistical techniques for lexical knowledge acquisition was first proposed by Bindi et al. (1991). In particular, that study illustrates the use of Multidimensional Scaling (MDS) to identify the relations among semantically related word types (e.g., *piccolo, corto, breve*) and their collocates (or 'word mates'). The input data for this approach are mutual information indexes computed over the domain of an entire corpus. Separate association indexes are computed for each pair of target word types (e.g., *piccolo* and *corto*) and for all of the word types in relation to various 'word-mates' (e.g., *piccolo* with *bambino*). This entire matrix is then analyzed by MDS to cluster word types and word-mates along a few underlying dimensions, providing a geometric representation of the semantic domain in question (in this case, 'smallness'). Word types are clustered together to the extent that they are associated with the same word-mates. The geometric distance between a word-mate and its word type reflects the strength of the relevant association index, and word-mates are clustered together to the extent that they are associated with the same word types (see Bindi et al. 1991).

Factor analysis differs from MDS in that the input data are computed over the domain of individual texts, rather than over the domain of the entire corpus. Further, the application of factor analysis here is to identify basic word senses and uses associated with a single word rather than the relations among a set of semantically related words. This statistical technique identifies the groupings of collocations that tend to co-occur frequently in texts. If we assume that texts are topically coherent, it follows that groupings of collocations that co-occur frequently in texts will often reflect different underlying word senses.

This approach is illustrated here through analyses of two words—*certain* and *right*—in an 11.66 million-word subsample of the Longman/Lancaster corpus. This corpus is designed to represent a wide range of text varieties (see Summers 1991), including ten major topical categories (e.g., natural science, social science, fiction), three mediums (books, periodicals, and ephemera), and three levels (technical/literary, lay/middle, and popular).

The first step in the analysis is to identify the major collocational patterns for the target word. In the present analysis, this was done simply by computing the frequency of all collocational pairs over the entire corpus; collocations that occurred more than 30 times were considered to be important word associations and thus included in the subsequent analyses.[3] For *certain*, 20 left collocates and 14 right collocates met this

2 Church et al. (1991, pp. 150–155) show how word association measures computed over the domain of discourse units (rather than bigrams or SVO triples) can be used to identify topical domains for information retrieval purposes.

3 There are more sophisticated approaches that could be used to identify the major collocational associations. First, mutual information indexes could be computed to identify strong word associations

criterion; for *right*, there were 24 left collocates and 53 right collocates meeting the criterion.

The next step was to count the frequency of each collocation pair in each text of the corpus. Only texts longer than 20,000 words were included at this stage, to insure adequate representation of the relevant co-occurring collocations. Two hundred forty-six texts from the subcorpus met this condition.

Factor analysis was then used to identify the groupings of collocational pairs that tended to co-occur in texts. Factor analysis builds upon pairwise correlations among the variables to identify a reduced set of underlying constructs, or 'factors'.[4] Each variable has some correlation, or 'loading,' with each factor, but only the larger loadings are important in interpreting the underlying constructs.

In the present case, the variables are the frequency counts for each collocational pair. The factors thus represent the collocational pairs that tend to co-occur frequently in texts. Given the following assumptions, it was hypothesized that the factor groupings of collocational pairs would represent different underlying word senses; the required assumptions are: 1) that each collocational pair tends to have a strong relation to a single word sense, and 2) that texts are topically coherent, and that words will therefore tend to be used in a single sense throughout the domain of a text. To the extent that these assumptions are accurate, collocational pairs should co-occur in texts as reflections of the same underlying word sense. The pilot analyses presented in the following section indicate that these assumptions do commonly hold in natural discourse.

3. Co-occurrence Patterns among Collocations for *certain* and *right*

The factor analyses for the collocates of *certain* and *right* are summarized in Tables 1 and 2 respectively.[5] The tables present the factor loadings for each collocational pair on each factor. Loadings can range from 0.0 to plus or minus 1.0. A loading of 0.0

---

that are not necessarily frequent in their overall occurrence. In addition, word associations at a distance should be included.

4 See Biber (1988) for a fuller discussion of factor analysis and its application to the computational analysis of linguistic variation. There are three main corpus design considerations required by this use of factor analysis. First, the analysis requires long, connected texts, to provide ample opportunity for the (co-)occurrence of a range of collocational pairs; the analysis here excludes all texts shorter than 20,000 words. Shorter texts would often contain few tokens of a few collocational pairs and would thus not adequately represent the correlations among collocations. Composite texts are not acceptable because they violate the assumption of topical coherence. Second, the analysis requires a large number of texts. A general rule of thumb for factor analysis is that there should be five times as many texts as variables (Gorsuch 1983). Thus, there should be a minimum of 240 texts included for the final factor analysis of *right*, which is based on 48 collocational pairs. (The final factor analysis of *certain* is based on 25 collocational pairs and would thus require only 150 texts.) The sub-corpus used here, with 246 coherent texts longer than 20,000 words, meets these criteria. Finally, the corpus used for analysis must represent the full range of variability with respect to the collocational patterns. Corpora that are restricted topically or restricted to a few registers are not adequate for analyses of this type, because the reduced variability results in reduced correlations, which in turn result in a skewed and unreliable factorial structure (cf. Biber 1990).

5 I used a common factor analysis with a Promax rotation. Scree plots of the eigenvalues and consideration of the two-, three-, and four-factor solutions indicated that the three-factor solution was the most adequate for the analysis of *certain*, while the four-factor solution was the most adequate for the analysis of *right*. Variables that had communalities $< .10$ were dropped from the final analysis; 25 collocational pairs were included in the final analysis of *certain*, and 48 pairs in the analysis of *right*.

In the analysis of *certain*, the first factor accounts for 21.4% of the shared variance, while the three factors together account for 39.4% of the shared variance. Factors 1 and 2 have a correlation of .30, while the other inter-factor correlations are negligible. In the analysis of *right*, the first factor accounts for 22.1% of the shared variance, while the four factors together account for 51.3% of the shared variance; the only appreciable inter-factor correlation is between Factors 2 and 3 (.43).

Table 1
Factorial structure of the collocates of *certain*.

		FACTOR1	FACTOR2	FACTOR3
Major Factor 1 Collocation				
in	certain	0.74077	0.07110	−0.05468
certain	other	0.68352	−0.09431	−0.05035
of	certain	0.61015	0.32266	−0.05660
and	certain	0.58373	−0.12996	−0.11671
certain	of	0.56773	−0.10960	0.22018
.	certain	0.49718	0.09231	−0.12760
,	certain	0.46448	0.10248	0.05560
there BE	certain	0.32284	−0.03747	0.07349
certain	type(s)	0.31556	0.06191	−0.10855
on	certain	0.29994	0.27142	−0.01401
with	certain	0.29908	0.24981	−0.07999
that	certain	0.27814	0.18763	0.01827
Major Factor 2 Collocations				
certain	extent	−0.21295	0.91065	0.01578
certain	aspect(s)	−0.08321	0.90127	0.01704
to	certain	0.15364	0.71625	0.00352
under	certain	0.02432	0.44979	−0.02374
for	certain	0.40779	0.44289	0.05722
a	certain	0.01796	0.35752	0.01724
by	certain	0.22450	0.33351	0.02019
Major Factor 3 Collocations				
certain	that	0.13828	0.00372	0.87341
certain	,	−0.11735	0.03704	0.46955
it BE	certain	−0.04318	0.00884	0.42112
make/made	certain	0.02137	0.00068	0.32323
I/we BE	certain	−0.08456	−0.06629	0.29203
quite	certain	−0.08470	0.04002	0.23607

shows that the variable has no relation to the pool of shared variance accounted for by the factor, while a loading of 1.0 represents a perfect correlation.

Each factor comprises a number of collocational pairs with relatively large loadings, while the remaining collocations have small or negligible associations. Tables 1 and 2 are organized so that the collocations having large loadings on each of the factors are grouped together.

Consider first the factorial structure for *certain*, presented in Table 1. The first 12 collocations listed on the table have large loadings on Factor 1, with generally small loadings on the other two factors (e.g., *in certain* has a loading of .74 on Factor 1, but loadings near 0.0 on Factors 2 and 3). The second group of 7 collocations have the largest loadings on Factor 2; while the last group of 6 collocations have large loadings on Factor 3.[6]

For the purposes of interpretation, each factor can be considered as comprising

6 The collocation *for certain* has a noteworthy loading of .41 on Factor 1 in addition to its loading of .44 on Factor 2.

Table 2
Factorial structure of the collocates of *right*.

		FACTOR1	FACTOR2	FACTOR3	FACTOR4
Factor 1 Collocations					
right	hemisphere	0.98345	−0.01528	−0.01434	−0.00717
right	sided	0.98343	−0.01487	−0.01405	−0.00737
right	hander(s)	0.98343	−0.01487	−0.01405	−0.00737
right	ear	0.98304	−0.01521	−0.00534	−0.00933
and	right	0.97245	0.00920	−0.01713	−0.00284
of	right	0.94451	−0.02987	−0.01721	−0.00597
the	right	0.93976	0.01421	−0.01173	−0.01295
right	side	0.87188	0.07561	−0.02721	0.02523
a	right	0.84510	0.00909	0.03104	−0.00871
or	right	0.80412	−0.01508	−0.03749	0.01806
right	hand	0.79835	0.04828	0.04005	−0.02339
to	right	0.60848	−0.03983	−0.03164	0.00806
right	eye	0.52883	−0.04162	0.14620	−0.00923
that	right	0.42500	−0.02100	0.31563	−0.00699
right	and	0.31707	0.04224	−0.03306	0.21127
right	of	0.29633	−0.06270	−0.10545	0.00798
right	as	0.24007	0.06985	0.01180	0.21477
Factor 2 Collocations					
go/went	right	−0.00434	0.72062	−0.05389	−0.14694
right	there	0.00397	0.65902	−0.00984	0.02282
right	back	0.00094	0.59946	−0.16701	−0.01122
right	now	0.01400	0.57063	0.06385	0.16459
right	out	0.00941	0.56666	−0.03302	0.03100
right	on	−0.00409	0.54770	−0.00731	0.02745
right	away	−0.02119	0.48780	0.00373	−0.08896
me/you	right	−0.01704	0.45494	0.11834	0.03465
right	here	−0.00095	0.45346	−0.05408	0.23709
right	for	−0.02277	0.44136	0.13048	−0.11498
right	up	−0.01647	0.43549	0.12772	−0.00086
right	in	0.02250	0.42069	−0.02334	0.11543
BE	right	0.00832	0.40980	0.16436	0.05745
it	right	0.00461	0.39168	−0.08280	0.08606
right	with	−0.00674	0.31426	0.12281	−0.14323
Factor 3 Collocations					
right	.	0.05717	0.10963	0.90268	0.02683
right	,	−0.00212	0.03678	0.88339	0.04162
all	right	−0.01675	0.09288	0.86456	−0.08946
that's	right	−0.00408	0.08106	0.68647	−0.14901
right	then	−0.00242	−0.07914	0.66223	0.03945
not	right	−0.00786	−0.15046	0.62236	−0.03184
.	right	0.01465	0.09568	0.56646	0.20960
right	a	−0.00220	−0.05189	0.35346	0.15038
you're	right	−0.01759	0.17132	0.19201	0.05199

Table 2
Continued.

		FACTOR1	FACTOR2	FACTOR3	FACTOR4
Factor 4 Collocations					
,	right	0.03424	0.02408	−0.12703	0.99315
right	you	−0.00214	−0.06841	0.10162	0.86214
right	so	0.02173	−0.05118	0.08467	0.77789
right	she/he/they	−0.00184	−0.07307	0.09301	0.59287
right	I/we	−0.00618	0.08199	0.05781	0.57046
right	it	0.01641	0.11691	−0.09122	0.37539
right	from	−0.01475	−0.00082	−0.03191	0.34669

only those features with large loadings, so that each factor represents a separate grouping of collocations that tend to co-occur frequently in texts. Factor 3 is the easiest to interpret: all six of the collocational pairs grouped on this factor represent contexts where *certain* functions to mark certainty. In all of these cases, *certain* is a predicative adjective, often taking a *that* complement clause (e.g., *I am quite certain that...*).

In contrast, *certain* does not have the sense of certainty in any of the collocational pairs grouped on Factors 1 and 2; rather the collocations grouped on these factors function to identify an unspecified (and perhaps unknown) subset of some larger group (e.g., *in certain cases, a certain person*).[7] The difference between Factors 1 and 2 is less obvious. An examination of the concordance listings for the associated collocations, however, shows that there is an important, systematic difference between the two factors: the collocations on Factor 1 tend to modify concrete, physical or tangible referents, while the collocations on Factor 2 tend to modify abstract referents. Thus, consider the examples from both factors listed in Table 3.

Factor 1 collocations tend to be more concrete, often characterizing physical objects (e.g., *towns, adults, trees, raw vegetables*). In contrast, Factor 2 collocations modify more abstract referents, such as *basic truths, general values,* and *assumptions.* The collocation *for certain* has a loading of about .40 on both Factor 1 and Factor 2, and it shows both kinds of pattern; for example, it modifies *users* and *speakers,* similar to other Factor 1 collocations, and it modifies *responsibilities, symptoms,* and *deficiencies,* similar to other Factor 2 collocations.

The factorial structure for *right,* presented in Table 2, similarly shows a highly systematic patterning among collocations. Factor 1 represents the positional or directional use of *right;* in fact, many of these collocations refer to body parts on the *right* side.

Factor 2 is equally transparent but represents a word sense that does not receive much attention in most dictionaries: *right* marking 'immediately,' 'directly,' or 'exactly.' There are a large number of collocational pairs having this sense (e.g., *right there, right back, right now*), and Factor 2 shows that there is a strong tendency for these collocations to co-occur in texts. On first consideration, the collocation *go/went right* seems to be an exception to the general pattern underlying Factor 2 (representing the Factor 1 sense of 'to the right' instead). Examination of the concordance listings for this collocation shows, however, that it almost always occurs with the Factor 2 sense, as in *go right through, go right up, he went right away.*

7 The collocational pair *for certain* does sometimes convey a sense of certainty, as in *it isn't known for certain whether...*

Table 3
Example contexts for major collocations on Factors 1 and 2 in the analysis of *certain*.

Factor 1 collocations

in certain: *ways, cases, instances, towns, districts, directions, areas, cultures, companies*

certain other: *external parasites, parts of the body, adults, statements, mediterranean cultures, crisp greens*

of certain: *monkeys, infants, theologians, particles, sovereigns, people, details, trees, individuals*

and certain: *kinds of cats, kinds of laughing, spices, bits of information, compositions, broiler growers, raw vegetables*

for certain: *users, kinds of addressees, speakers, materials*

Factor 2 collocations

(to) a certain extent
(for) certain aspects

to certain: *deficiencies, stimuli, nonmaterial aspects, very strong drives, basic truths, defects, general values, simplifying principles*

under certain: *assumptions, (economic) conditions*

for certain: *responsibilities, symptoms, deficiencies, types*

Factor 3 represents a grouping of collocations having the general sense of 'ok' or 'correct.' *All right* is commonly used to mark agreement or to mark a discourse juncture. *Right* can occur by itself with these same functions (hence the collocations of *right* with preceding and following punctuation). The collocation *that's right* also typically marks agreement to a previous assertion. *Right then* appears to represent a use of the Factor 2 pattern meaning 'immediately,' but this collocation also frequently occurs as a response in the context *all right then*. The collocation *not right* indicates a lack of correctness or normalcy, as in *all was not right with the mirror* and *something was not right close-up*.

Finally, Factor 4 seems to represent a stylistically marked use of *right* at the end of a clause, with no intervening punctuation before the following clause. Collocations of this type commonly have *all right, right,* or *that's right* preceding a clause, as in *all right you found the...*, *all right I pushed...*, *that's right I think, right away they walk toward...* The underpinnings of this factor need further investigation.

4. Conclusion

The two pilot analyses presented here indicate that this approach can be a useful tool for the semi-automatic identification of underlying word senses and uses. Further, both analyses produced unanticipated but systematic results, indicating that this approach can provide a useful complementary perspective to traditional lexicographic methods.

These analyses could be extended in several ways. First, statistics such as the mutual information index could be used to help identify the set of relevant collocations to be included in the factor analysis. Second, the corpus could be pre-processed by a

grammatical tagger, making the collocations sensitive to grammatical category. Finally, lexical associations at a distance should be included in the analysis. In this regard, the analyses here have been restricted: they consider only collocations of adjacent words, with no regard for grammatical category, and with the input data identified simply on the basis of absolute frequency. However, even with these restrictions, factor analysis appears to be a powerful tool for identifying underlying patterns among collocations, reflecting some of the major senses and uses of a word.

References

Biber, Douglas (1988). *Variation across Speech and Writing*. Cambridge University Press.

Biber, Douglas (1990). "Methodological issues regarding corpus-based analyses of linguistic variation." *Literary and Linguistic Computing*, 5, 257–269.

Bindi, Remo; Calzolari, Nicoletta; Monachini, Monica; and Pirrelli, Vito (1991). "Lexical knowledge acquisition from textual corpora: A multivariate statistic approach as an integration to traditional methodologies." In *Proceedings, Seventh Annual Conference of the UW Centre for the New OED and Text Research*. Oxford, U.K.

Calzolari, Nicoletta, and Bindi, Remo (1990). "Acquisition of lexical information from a large textual Italian corpus." In *Proceedings, 13th International Conference on Computational Linguistics*. 54–59.

Church, Kenneth Ward, and Hanks, Patrick (1990). "Word association norms, mutual information, and lexicography." *Computational Linguistics*, 16, 22–29.

Church, Kenneth; Gale, William; Hanks, Patrick; and Hindle, Donald (1991). "Using statistics in lexical analysis." In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, edited by Uri Zernik, 115–164. Lawrence Erlbaum Associates.

Gorsuch, Richard L. (1983). *Factor analysis*. Lawrence Erlbaum Associates.

Sinclair, John (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Smadja, Frank (1991). "Macrocoding the lexicon with co-occurrence knowledge." In *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, edited by Uri Zernik, 165–189. Lawrence Erlbaum Associates.

Summers, Della (1991). "Longman/Lancaster English language corpus: Criteria and design." Technical Report, Longman.

Zernik, Uri (1991). "Introduction." In *Lexical Acquisition: Exploiting On-line Resources to Build a Lexicon*, edited by Uri Zernik, 1–26. Lawrence Erlbaum Associates.