Co-occurrence Patterns and Lexical Acquisition in Ancient Greek Texts

Jeffrey A. Rydberg-Cox
University of Missouri at Kansas City, Kansas City, MO, USA

Abstract

Many recent studies have demonstrated that the analysis of word co-occurrence patterns can be a valuable tool for the acquisition of lexical knowledge from unstructured texts. The success of these studies raises the question of whether these results can be replicated in other languages and, in particular, in ancient Greek texts. The adaptation of these techniques for the acquisition of lexical knowledge could provide the foundation for useful philological and lexicographical research tools and multi-lingual information retrieval applications. One method that has proven useful in the study of Greek texts is the use of co-occurrence data to calculate mutual information scores. The approaches used for English language texts, however, require some modification for use with ancient Greek. This paper will describe the methods and modifications required for the analysis of collocational data from unstructured ancient Greek texts so that they produce useful results.

1 Introduction

Many recent studies have demonstrated that the analysis of word co-occurrence patterns can be a valuable tool for the acquisition of lexical knowledge from unstructured texts.[1] The success of these studies raises the question of whether these results can be replicated in other languages and, in particular, in ancient Greek texts. The adaptation of these techniques for the acquisition of lexical knowledge could provide the foundation for useful philological and lexicographical research tools and multi-lingual information retrieval applications. The evaluation of these methods is also an important first step in the creation of a new Greek lexicon for an electronic environment. The practice of classical philology has almost entirely ignored the field of computational linguistics except when studying questions of authorship and dating. However, projects such as the COBUILD dictionary and grammar series and WordNet demonstrate the broader usefulness of these techniques and suggest that they should not be ignored in the study of classical philology. One method that has proven useful in the study of Greek texts is the use of co-occurrence data to calculate mutual information scores. The approaches

Correspondence:
Jeffrey A. Rydberg-Cox,
Department of English,
University of Missouri at Kansas City, Cockefair Hall, 106 Kansas City, MO 67110, USA.
E-mail:
jrydberg@perseus.tufts.edu

1 See Church and Hanks (1990), Church *et al.* (1991), Sinclair (1991), Smadja (1991), and Biber (1993).

used for English language texts, however, require some modification for use with ancient Greek. This paper will describe the methods and modifications required for the analysis of collocational data from unstructured ancient Greek texts so that they produce useful results.

2 The Testbed

The testbed for this research is the corpus of ancient Greek texts in the Perseus digital library. The Perseus digital library is a heterogeneous collection of texts and images pertaining to the Archaic and Classical Greek world, late Republican and early Imperial Rome, and the English Renaissance.[2] The texts are integrated with morphological analysis tools, student and advanced lexica, and sophisticated searching tools that allow users to find all of the inflected instantiations of a particular lexical form.[3] The current corpus of Greek texts contains more than four million words by thirty-three authors. Most of the texts were written in the fifth and fourth centuries BCE, with some written as late as the second century CE. These texts represent a wide range of genres, dialects, and styles, provide excellent coverage of an extremely well-studied period, and are large enough to yield significant results from computational methods.

3 The Approach

The program to calculate mutual information scores for Greek words traverses a text, collection of texts, or the entire Perseus corpus, and counts the words that appear within a window of five words to the left or right of each word in the corpus.[4] It also notes the frequency of individual words in the corpus. After obtaining these frequency scores, a mutual information score can be calculated for each collocational pair. A mutual information score is a ratio expressing the frequency with which two words appear in proximity relative to the frequency of each word occurring independently within the corpus. As this ratio increases, the two words become more likely to appear together. The equation used for this calculation is

$$I(x,y) = \log_2[P(x,y)/P(x)P(y)]$$

where $P(x,y)$ is the weighted frequency score for terms x and y normalized by the weighted score of the number of words in the corpus, $P(x)$ is the weighted frequency score of term x, and $P(y)$ is the weighted frequency score of term y, both also normalized by the number of terms in the corpus.

During the counting phase, the program consults the Perseus morphological analysis program and resolves each word to its lexical form. Approximately 20 per cent of the words in Perseus are lexically ambiguous in that they can be derived from two or more dictionary headwords. The Perseus morphological analyser currently only returns the different possible lexical forms and does not make any attempt to resolve this ambiguity.[5]

2 The Greek materials have been published on several CD-ROMs that are available from Yale University Press. All of the materials in the Perseus Digital Library are freely available on the World Wide Web at www.perseus.tufts.edu

3 The Greek parser is described in Crane (1991).

4 A window of five words is used because the primary goal of this program is to obtain a picture of word senses and co-occurrence patterns rather than identifying fixed phrases and idioms. A smaller window would facilitate the latter process.

5 Resolution of morphological, lexical, and word sense ambiguity is, however, an active part of the research agenda at the Perseus Project.

Two different approaches were tried to account for this lexical ambiguity. In the first iteration of this program, two notations were made for each word; lexically ambiguous words were noted in a table of the maximum possible collocations whereas non-ambiguous words were tallied in tables of both the minimum and maximum possible collocations. With this approach, two word frequency tables for each individual headword and the entire corpus were also created. These tables were then used to calculate a minimum and maximum mutual information score for each word pair. Although this approach produced acceptable results in some cases, it also introduced several problems. In many cases, relatively common Greek words are lexically ambiguous with relatively rare words. For example, many forms of the common Greek verb *echô*—meaning 'to have'—are lexically ambiguous with the less common Greek noun *echis*—meaning 'viper'. When using the two-score system, this type of rare word is over-represented in the maximum collocation tables whereas very common words are excluded from the minimum collocation tables. Both situations added obvious inaccuracies to the results.

To solve this problem, a simple weighting scheme was developed to replace the minimum and maximum score system in the second iteration of this program. Local weights were assigned to each inflected form based on their level of lexical ambiguity. The local weight was calculated as the reciprocal of the number of lexical forms from which a word could be derived. If, for example, an inflected form could be derived from four dictionary headwords, each of those headwords would receive a weight of 0.25. When that particular inflected form was encountered, the local weight was added to the frequency score for each possible headword. Weighted frequency scores for word pairs are calculated as the median of the weighted score of both words. This process reduced the prominence of less common words that share forms with more common words while more accurately reflecting the relative frequency of the more common words. Although even more accurate results could be obtained with a tagged corpus or more sophisticated stochastic disambiguation methods, this simple and computationally inexpensive weighting scheme significantly reduced the number of 'false positives' for common words in the collocation results.

Although this approach accounts for the problems of lexical ambiguity, there are also problems with the mutual information ratio itself that must be addressed. This ratio is problematic because it gives low-frequency collocations too much prominence; the highest mutual information scores appear for word pairs that do not appear very often within the corpus.[6] For this reason, two modifications to the mutual information equation are also necessary. First, following the approach of Church and Hanks (1990), the mutual information score was calculated only for pairs with a weighted frequency score of five or more. Second, to further correct for the tendency of mutual information scores to overemphasize low-frequency collocations, the ratio is multiplied by the log of the weighted pair frequency score.

6 See Dunning (1993) and Manning and Schütze (1999).

4 Problems Specific to Ancient Greek

Despite the success of collocational approaches for the acquisition of lexical knowledge from unstructured English texts, it is not immediately obvious that these techniques would apply to texts written in ancient Greek. Three specific features of the Greek language pose potential problems for this type of analysis and require modifications to the approach used for English language texts. First, Greek makes much more extensive use of function words than English. The most frequent word in any Greek text will almost always be the definite article and, following Zipf's law, it will usually occur twice as often as the next most frequent word in the corpus. This problem is compounded by the prevalence of particles in Greek texts. Greek particles are short words that provide emphasis, mark the tone of a sentence, or indicate the relationship of a sentence to the preceding and following sentences. These particles are also among the most common words in any corpus of ancient Greek. Further, authors frequently use two or more particles consecutively for some rhetorical effect. The relative frequency of these function words hinders the process of lexical acquisition by filling the collocational tables with function words rather than the content words that will be more useful and interesting for philological inquiry, lexicography, and information retrieval applications.[7] For this reason, a stop list is employed during the initial traverse of the texts. If a definite article or a Greek particle is encountered, it is ignored and the window is expanded by one word.

Second, Greek word order is far less structured than English word order. Ancient Greek is a highly inflected language; grammatical relationships such as subject, direct and indirect objects, adjectives, and nouns are expressed by gender, number, and case endings of noun and adjectives rather than by their position in a sentence. For example, the phrases *ho kakos anthrôpos*, *ho anthrôpos kakos*, and *anthrôpos ho kakos* can all be used to express the phrase 'the bad man'. This problem is further compounded because Greek authors consciously varied their word order for reasons of style, emphasis, and balance. In fact, one common style, known as the periodic style, allows authors to place the main verb as the last word of a sentence. Authors who write in this style also regularly include at least some subordinate clauses for rhetorical effect in these sentences. The practical consequence of this style is that these sentences can be very long. This problem is addressed in two ways. First, both right and left collocates are counted for every node word. Because of the relatively free nature of Greek word order, a focus on one direction produces incomplete rather than interesting results for the process of lexical acquisition.[8] Second, to account for the problem of long sentences, the window used in this implementation is relatively large. Approaches that examine the immediate right collocate of a word or slightly larger windows of two or three words do not produce interesting results in ancient Greek texts.[9] This problem is resolved by expanding the window within which words are considered. Although the base window in this implementation is five words, in practice, the above-described use

7 Despite their inappropriateness for this application, Greek particles are, in themselves, an interesting object of study. One of the masterworks of classical philology in the twentieth century is a study of these particles (Denniston, 1959). Further, Burrows (1987) has demonstrated that this sort of function word can benefit from computational analysis. A study of co-occurrence patterns of various particles with different words could prove to be an important tool for lexical and word sense disambiguation.

8 Although the equal treatment of left and right collocates is best suited for lexical acquisition, collocation patterns in one direction would be useful for more specialized studies of Greek word order such as Dover (1968) or Dik (1995).

9 See Sinclair (1991, pp. 67–79) and Biber *et al.* (1998, pp. 43–51) for studies of immediate right collocates. Biber *et al.* (1998, pp. 51–3) examined the collocations of the word 'large' within windows of two and three words.

of a stop list renders the window much larger, approaching ten words in most cases.[10]

The third modification to the approach used for English texts is also a product of the highly inflected nature of Greek. Many studies of collocations in English texts examine inflected rather than lexical forms. Using inflected forms is problematic in ancient Greek texts because very few of these forms appear together with sufficient frequency to produce a significant mutual information score. The 4.2 million word Perseus corpus consists of approximately 280,000 unique inflected forms; less than 2 per cent of these inflected forms appear in combination with each other five or more times. Although these word pairs would be extremely interesting for a more narrowly defined study of idioms or fixed phrases, or for a lexicographer writing a dictionary entry for a word that happened to be on this list, the results based on lexical forms are much more broadly interesting and applicable to a much wider range of words. Further, using lexical forms rather than inflected forms allows for the division of the Perseus corpus into smaller sub-corpora based on genre or style while still producing significant and interesting results. For these two reasons, the program has been fully implemented using lexical rather than inflected forms.

5 Results

This approach, when applied to ancient Greek texts, yields very interesting results that can provide an extremely powerful starting point for traditional philological research. Even by themselves, the mutual information scores can be a useful tool for students of Greek literature because they allow readers to undertake the sort of research suggested by Firth on a scale not previously possible. These collocation data become even more valuable when integrated with both the primary texts on which they are based and also with other lexical reference works. The program described here has been used to calculate a mutual information score for each Greek word in the Perseus corpus. These data have been integrated with Perseus' electronic Greek–English lexicon.[11] If a user looks up a word in the Greek lexicon, a table showing the five most common collocates of that word appears at the head of each dictionary entry. This table contains links to the dictionary entries for each collocate and also a link to a more detailed listing of the collocation data. This more detailed entry shows every collocational pair that occurs with a given node word five or more times, the mutual information score for each pair, and a link to an information retrieval system to display the texts in which the word pairs appear.[12] Collocation data have also been calculated for several sub-corpora of texts representing different styles and genres such as rhetoric, prose, tragedy, and poetry. If a user looks up a word while reading a text in one of these sub-corpora, the table that appears in the dictionary entry will show the five most common collocates for each applicable sub-corpus in addition to the collocates for the complete collection of Perseus Greek texts.

10 The collocation window is not, however, allowed to extend beyond a sentence boundary.

11 The integrated electronic reading environment of the Perseus digital library is described in Crane (1998).

12 Documentation and sample links for these tools can be seen at http://www.perseus.tufts.edu/PR/colloc.ann.html Documentation and sample links for the information retrieval system can be seen at www.perseus.tufts.edu/PR/search.ann.html

Integrating these collocation data with the electronic lexicon allows readers to quickly obtain a broad sense of the 'company that words are keeping' while also providing a rough guide to possible idioms and common phrases. One clear example of the usefulness of these tables for determining the broad senses of a word can be seen in the collocation data for the word meaning silver or a silver coin, *argurion*. The collocational pairs with the highest mutual information scores include nouns for different types of coins and verbs denoting the giving, taking, and lending of money.[13]

Table 1 Words that regularly appear with *argurion* in the Perseus Greek corpus (n = 4.2 million)

Word	Score	Word	Score
1. talanton (a weight or sum of money)	89.75	2. didômi (to give)	84.98
3. lambanô (to take or seize)	81.47	4. apodidômi (to give away or sell)	80.04
5. chrusion (a piece of gold, anything made of gold)	78.86	6. daneizô (to lend or borrow money)	78.22
7. mna (a type of money)	77.61	8. drachmê (a silver coin)	71.87

The value of these tables for drawing out the primary senses of a word can also be seen in the mutual information scores of the words that co-occur with the verb 'to sacrifice', *thuô*. The words with the largest mutual information scores describe the personnel, gods, and objects associated with the process of sacrifice.[14]

Table 2 Words that regularly appear with thuô in the Perseus Greek corpus (n = 4.2 million)

Word	Score	Word	Score
1. theos (god or goddess)	89.73	2. bômos (altar)	81.59
3. hieros (holy, consecrated)	79.49	4. thusia (offering or sacrifice)	78.82
5. hieron (a holy place or a temple)	77.75	6. Zeus	76.12

It is also possible to use these collocation tables as a rough guide to common phrases and idioms in Perseus Greek texts.[15] Two examples will illustrate the value of these tables for this purpose. First, throughout Greek rhetorical works delivered in Athenian law courts, speakers frequently address the jury as 'men of Athens'—*andres Athênaioi*. This common phrase is reflected in the collocation data; the word *Athênaios* appears as the most common collocate of the word *anêr* in the corpus of Greek rhetoric.[16] Similarly, Greek prose authors commonly use the phrase *kalos kai agathos* to describe a good and noble man. The collocation data also point to this common expression; the word *agathos* appears as the collocate of *kalos* with the highest mutual information score.[17]

6 Evaluation

The question of how to evaluate the results of this program remains. Although it is easy for a philologist to claim that the above-described

13 The dictionary entry containing this table can be found at www.perseus.tufts.edu/cgi-bin/lexindex?lookup=a)rgu/rion

14 See http://www.perseus.tufts.edu/cgi-bin/lexindex?lookup=qu/w

15 Although, as noted above, tweaking a few parameters in my program would probably result in better tools for the specific task of studying fixed phrases.

16 The dictionary and collocation table for *anêr* is located at www.perseus.tufts.edu/cgi-bin/lexindex?lookup=a)nh/r&author=dem

17 The dictionary and collocation table for *kalos* is located at www.perseus.tufts.edu/cgi-bin/lexindex?lookup=kalo/s

collocates are significant, this is a highly subjective evaluation process. The field of classical studies lacks established resources such as Roget's Thesaurus or WordNet against which these results can be checked. Three methods, however, are available for testing these results. First, by making the collocational data for every word in the Perseus corpus available to the users of the Perseus digital library, the results are subjected to a broad form of peer review.[18] Our continuing evaluation of the Perseus web site, including the analysis of server logs and interactions with users, will also help determine the usefulness of these data.

Second, comparison of the results with existing, formalized lexical knowledge also provides a somewhat more objective means to evaluate these results. For example, ancient Greek has two ways of expressing the idea of 'neither . . . nor', *oute . . . oute* and *mête . . . mête*. If this program is producing correct results, these collocational pairs should have a high mutual information score. This expected result, in fact, appears in the co-occurrence data for the Perseus corpus; the raw mutual information score for *oute . . . oute* is 8.939 and it is 10.617 for *mête . . . mête*.

Finally, although there are no large-scale reference works against which co-occurrence tables can be checked, current Greek lexica contain some information about specific words that can be used to verify these results. For example, the entry in the Liddell, Scott, and Jones Greek–English Lexicon for the word *agraphos*—meaning 'unwritten'—reports that *agraphos* is frequently used in combination with the word *nomos* to express the idea of an unwritten custom, law, or tradition.[19] This observation is replicated in the mutual information table for this word; the word *nomos* appears as the most common collocate of *agraphos* and the adjusted mutual information score is 112.74. Similarly, in Greek poetry it is common to describe people and divinities with certain epithets. For example, the goddess Athena is frequently described as the bright-eyed goddess Athena—'*thea glaukôpis Athênê*'. The collocation data shows that *thea* and *Athênê* are the two words that most commonly co-occur with *glaukôpis*.[20] This combination of methods will allow the mutual information and co-occurrence data to be trusted both for traditional research and other information retrieval applications.

7 Conclusions

The analysis of co-occurrence data and mutual information scores appears to be a useful tool for the acquisition of lexical knowledge from unstructured texts written in ancient Greek. The addition of a stop-list, the expansion of the collocation window, and the use of lexical rather than inflected forms render a method that has proven successful in English texts a useful tool for the study of Greek texts as well. The results of this program have broad applicability as a tool for traditional philological and lexicographic research when integrated with the electronic reading environment of the Perseus digital library. The results can be checked by the review process entailed in broad public availability and

18 The Perseus web site currently delivers 600,000–700,000 pages per week during the academic year.

19 The dictionary entry and collocation information for *agraphos* is available at http://www.perseus.tufts.edu/cgi-bin/lexindex?lookup=a)/grafos

20 The dictionary entry and collocation information for *glaukôpis* is available at http://www.perseus.tufts.edu/cgi-bin/lexindex?lookup=glaukw=pis&author=hom

more rigorous comparison of the results with other formalized grammatical and lexical knowledge.

Although these results on their own are substantial, they also serve as the starting point for other explorations. More detailed recording of the median distance and the variance between co-occurring words can provide a more precise guide to fixed phrases and idioms within texts.[21] The use of the Perseus morphological analysis tools to tag parts of speech would also allow for more detailed study of the syntax and grammar of particular words.[22] Finally, co-occurrence data provide the basis for other applications such as the generation of an English–Greek thesaurus that would be useful for multi-lingual information retrieval applications and also for more traditional humanistic research.[23] If these other areas of corpus-linguistic study can be applied effectively to ancient Greek texts, they would provide important tools for philological research and, ultimately, form the basis for an extremely useful working environment for the practice of Greek lexicography.

References

Biber, D. (1993). Co-occurrence patterns among collocations: a tool for corpus-based lexical knowledge acquisition. *Computational Linguistics*, **19**(3): 531–8.

Biber, D., Conrad, S., and Reppen, R. (1998). *Corpus Linguistics: Investigating Language, Structure and Use*. Cambridge: Cambridge University Press.

Burrows, J. F. (1987). *Computation into Criticism: a Study of Jane Austen's Novels and an Experiment in Method*. Oxford: Clarendon Press.

Church, K. and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, **16**(1): 22–9.

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using statistics in lexical analysis. In Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum.

Crane, G. (1991). Generating and parsing Classical Greek. *Literary and Linguistic Computing*, **6**: 243–5.

Crane, G. (1998). New technologies for reading: the lexicon and the digital library. *Classical World*, **91**: 471–501.

Denniston, J. D. (1959). *The Greek Particles*. Oxford: Clarendon Press.

Dik, H. (1995). *Word Order in Ancient Greek: A Pragmatic Account of Word Order Variation in Herodotus. Amsterdam Studies in Classical Philology; Vol. 5*. Amsterdam: J. C. Gieben.

Dover, K. J. (1968). *Greek Word Order*. Cambridge: Cambridge University Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**: 61–74.

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston, MA: Kluwer Academic.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus Based Approach*. Oxford: Clarendon Press.

21 See Church and Hanks (1990), Smadja (1991), and Moon (1998).

22 Biber *et al.* (1998, pp. 84–105).

23 Srinivasan (1992) and Grefenstette (1994).

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

Smadja, F. (1991). Macrocoding the lexicon with co-occurrence knowledge. In Zernik, U. (ed.), *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, NJ: Lawrence Erlbaum.

Srinivasan, P. (1992). Thesaurus construction. In Frakes, W. and Baeza-Yates, R. (eds), *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, NJ: Prentice Hall.