Vocabulary Building in the Perseus Digital Library

Jeffrey A. Rydberg-Cox Anne Mahoney

February 6, 2002

1 Introduction

Vocabulary acquisition is a particularly vexed question for intermediate students of Greek and Latin. Cognitive studies of language acquisition suggest that a word must be encountered between 6 and 20 times before a student can be said to *know* that word (Nation [1990], p. 43-45; Parry [1997]; Coady [1997]; Hulstijn [1997]). At the same time, encountering these words on general vocabulary lists or flash cards of common words intended for rote memorization does not help; retention rates from lists that are not connected to reading are extremely low. Students must encounter new words in context in order to retain their meanings (Cf. Perry [1998] p. 111-112). Unfortunately, the usual distribution of words in texts makes it difficult if not impossible for students to learn vocabulary simply by reading. Vocabulary is distributed in literary texts according to a phenomenon known as Zipf's law. This law states that the most common words in any text will be "function words," like the definite article, pronouns, conjunctions, or common prepositions, while most words in a text will appear very few times and the vast majority will appear only once (Zipf [1935], Baayen [2001]). In the course of normal reading, therefore, students will not encounter a word in context often enough to add it to their active vocabulary. One solution to this problem is linking reading assignments to vocabulary lists so that students can efficiently study the words that they are encountering in their readings. In this paper, we will describe a new computational tool in the Perseus Digital Library designed to help students learn vocabulary by generating Latin and Greek word lists that are tailored to reading assignments.

Of course experienced teachers know that it takes more than just vocabulary to read a lan-

guage, and especially to read literary language, a point made strongly by Kitchell [2000]. In addition, familiar words can be used in unexpected ways, and obscure words can be crucial for the meaning of a passage (as argued by Bull [1948-1949]). Nonetheless, vocabulary is one part of language learning, and is all we focus on here.

2 The Perseus Vocabulary Tool and How It Works

The Perseus Digital Library (`http://www.perseus.tufts.edu`) now includes a vocabulary tool that can generate several different kinds of vocabulary lists for Greek or Latin texts. Whenever a text is displayed in the Perseus Digital Library, the library also offers a link to a basic vocabulary list of the words that appear in that text. Users can also create a list for more than one work using a separate interface that displays all the Greek or Latin works in a Perseus collection. You can also create vocabulary lists for smaller sections of texts that can be logically divided into smaller units such as the books of Virgil's *Aeneid* or Herodotus' *History*. Thus, a Greek survey course could create a list containing the vocabulary for Book 1 of the *Iliad*, Demosthenes' *Against Neaera*, and Aeschylus' *Agamemnon*. Likewise, an Advanced Placement Latin course might construct a vocabulary list for Cicero's *Pro Caelio* and selected poems of Catullus, or for the relevant sections of the *Aeneid* (See the sample in figure 1).

The custom vocabulary list interface allows readers to change the way that the list appears with several different sort, filtering, and output options. First, it is possible to sort the list either alphabetically or by word frequency. Sorting in alphabetical order produces a traditional word list, convenient for looking up words while reading the text. Sorting by frequency puts the most common words at the top of the list, making it easy for students to see the most basic words they need for a text.

Counting word frequencies for Greek and Latin texts is more complicated than it might seem. The current version of the Perseus morphological analyzer, described in Crane [1991], makes no attempt to disambiguate forms that can be derived from more than one lexicon entry, as, for example, the English word "flies" might come from the verb "to fly" or the noun "a fly" but the word "flew" is unambiguously a form of "to fly." Word forms that are ambiguous are included in the maximum count for each dictionary word they might belong to, while unambiguous forms are included in both the minimum and maximum counts. We also calculate a weighted

frequency that attempts to show whether the actual frequency count for a word would be closer to the minimum or maximum frequency score. Note that a form with a minimum weight of zero means that every instance of a word in a text is ambiguous and that the word may not actually appear in the text at all. For example, an English text about airplanes may contain forms of the verb "to fly" but no mention of the insect called "fly," yet "a fly" may appear in the vocabulary list because the form "flies" *might* have come from that word.

The tool also allows two different mechanisms for viewing the list: you can choose a table that will provide attractive output in a web browser, or a comma-delimited list that you can import into other software programs such as a spreadsheet or database. Finally, the vocabulary tool allows you to select the percentage of the words in a document that you want to include in your list. As with the sort orders, the different percentages are useful for different purposes. Since the vast majority of words in any text appear only once, vocabulary lists showing all words in a document can be quite long. A complete list is precisely what is wanted for comprehensive review or a "mini-lexicon" for a selection of works. If, on the other hand, you want to give students a list that contains the essential vocabulary for the selected texts, you can include only the words that account for a higher percentage of the words in the text.

Consider Ovid's *Metamorphoses*. The text is over 78,000 words long, but Ovid uses only 8,789 different words. Of those, 3,644 — almost half — appear only once. The most frequent words in the *Metamorphoses* are *et, sum, in,* and *qui*, appearing more than 1,000 times each. Half of the 78,000 words in the poem are forms of only 321 different words. In other words, a student who knows those 321 words will know, on average, half the words on a page of the *Metamorphoses*. Three quarters of the total are forms of 1,200 different words. To get to 90%, you need 3,000 words. For 95%, 4,575 words suffice. The 95-percent level is significant because a student who knows 95% of the words in a text can usually figure out most of the rest from context (Nation and Coady [1988], Laufer [1997]). Hence, although Ovid's vocabulary is large, there is no need to learn every single one of those 8,789 words before starting to read.

For each word in the list, the vocabulary tool also calculates what we call a "key term score." (It's calculated using a standard metric from information retrieval and computational linguistics, known as $tf \times idf$. This calculation is described in Salton and Buckley [1988], Salton [1989], Singhal et al. [1996].) The key term score provides a guide to words that appear relatively frequently in the works on the vocabulary list but relatively infrequently in the rest

Latin Vocabulary List For:
Cicero, For Marcus Caelius.

Words:	8473	Unique Words:	2248	Vocabulary Density:	3.769
Possible Key Terms:	59			Words occurring only once:	1318

Count	Word	Max. Freq.	Min. Freq.	Weighted Freq. *	Key Term Score	Running Total of Wgt. Freq.	% of Total	Running % of Total	Short Definition
	qui	381	61	213.92	0.00	213.92	2.52	2.52	who, which, what
	sum	308	119	212.75	0.00	426.67	2.51	5.04	to be, exist, live
	in	178	178	178.00	0.00	604.67	2.10	7.14	unequal
	non	173	171	171.67	0.00	776.33	2.03	9.16	not, by no means, not at all
	quis	296	50	165.92	0.00	942.25	1.96	11.12	plur
	et	151	151	151.00	0.00	1093.25	1.78	12.90	also, too, besides, moreover, likewise, as well, even: Ph
	hic	165	127	146.00	0.00	1239.25	1.72	14.63	this, this here
	si	119	93	102.00	0.00	1341.25	1.20	15.83	if, when, inasmuch as, since
	cum	101	98	99.50	0.00	1440.75	1.17	17.00	when, as, while
10	edo	184	0	92.00	0.00	1532.75	1.09	18.09	
	ab	82	82	82.00	0.00	1614.75	0.97	19.06	out of, from
	atque	111	49	80.00	0.00	1694.75	0.94	20.00	and, as well as, together with
	ego	81	76	77.00	0.00	1771.75	0.91	20.91	I, me, we, us
	dico	84	69	76.08	0.00	1847.83	0.90	21.81	to proclaim, make known
	ut	75	75	75.00	0.00	1922.83	0.89	22.69	where, when, as
	tu	83	64	73.50	0.00	1996.33	0.87	23.56	thou, you
	is	103	46	72.83	0.00	2069.17	0.86	24.42	he, she, it, the one mentioned
	ad	69	69	69.00	0.00	2138.17	0.81	25.24	to, toward

Figure 1: Sample Vocabulary List for Cicero's *Pro Caelio*

of the collection in the Perseus Digital Library. Words with a high key term score provide an initial guide to important people, places, and concepts in your selection of texts. Frequently appearing words that provide less guidance about the contents of your selection will have a low key term score, and the least important words will have a score of zero. Very common words like *sum* or *ille* in Latin, *eimi* or *outos* in Greek, will always have a key term score of zero. Proper names, on the other hand, often have relatively high key term scores because they are often the most distinctive words in a text. Another way to look at it is that words with a non-zero key term score are the most useful words to learn before starting to read this text: they are the ones that an intermediate-level student might not already know, but that are frequent enough in this text to be worth learning. Although only the first five or ten "key terms" tell you about the content of the text, the complete key term list gives you an overview of the likely new vocabulary.

For example, let's look at the words with a high key word scores for two documents, Lysias' *On the Murder of Eratosthenes* and Book 21 of the *Odyssey*. The top ten key words for *On the Murder of Eratosthenes* include the name Eratosthenes and words for adultery, a servant woman, a child, a door, and several words for entering a house. Likewise, the top key words for Book 21 of the *Odyssey* include Antinous, Odysseus, Telemachus, and nouns and verbs associated with stretching and stringing a bow. The key words for these two document do not, of course, capture all of the nuances of the actions being described, but they do provide a useful overview of elements and that are potentially important and unfamiliar vocabulary as you read the texts.

Finally, the vocabulary lists also include short definitions that have been automatically extracted from the *Intermediate Liddell and Scott Greek Lexicon* and Lewis' *Elementary Latin Dictionary* (Rydberg-Cox [Forthcoming 2001]). Because this definition is the one listed first in the dictionary entry for each word, the definition provided for words with multiple senses may not be entirely correct for the works that you have selected, and words that are not in these medium-sized dictionaries won't have definitions at all. The vocabulary tool has two different facilities to address these shortcomings of the automatically-extracted definitions. The words in the HTML vocabulary list are linked to the Perseus Word Study Tool (described further in Mahoney [2001]), from which you can look up the full definition in either the intermediate or unabridged dictionaries in the Perseus Digital Library. The vocabulary tool also provides the

facility to eliminate the short definitions from the vocabulary list. It is also possible to include or suppress other columns in the vocabulary list, including the frequencies, the percentages, and the key term scores. In fact, the only column that is not optional is the words themselves.

2.1 Other Information with the List

In addition to the vocabulary list itself, the vocabulary tool provides three quantitative measures designed to provide you with a broad sense of how complex the vocabulary is in your selection of texts. The simplest statistic is a count of the total number of words in your selection of documents. Second, the tool calculates the number of unique words in your document or, in other words, the number of distinct dictionary words used in the text or collection of texts. The third number calculated by the system is a "vocabulary density" score. This score is the ratio of the total number of words in the document to the number of unique words in the document. A work with more complex vocabulary will have more unique words while a work with simpler vocabulary will have fewer unique words. The vocabulary density ratio provides a normalized way to view this information. Higher scores, in general, mean easier texts while lower scores usually point to more difficult texts. It is important to note, however, that vocabulary density scores can only be used for extremely rough comparison among works because the score is highly dependent on the length of the vocabulary list. The density score for a portion of a larger work, moreover, will almost always be lower than for the work as a whole (See Tweedie and Baayen [1998], Yule [1938], and Yule [1944]). Another way to think about this score is that it is a rough expression of the number of words on average that you will encounter between new words.

Compare, for example, the word counts and vocabulary density scores for Aeschylus' *Oresteia* and Xenophon's *Anabasis*. The *Oresteia* contains 18,934 words and 6,974 unique words with a vocabulary density score of 2.715. This means that, on average, one out of every three words that a reader encounters will be new. On the other hand, Xenophon's *Anabasis* contains 57,193 words with 4,358 unique words, for a vocabulary density score of 13.124. The higher vocabulary density score suggests a much simpler vocabulary; on average only one in every thirteen words will be new. In fact, the *Anabasis* is three times longer than the *Oresteia* but it contains only about two-thirds as many unique words. Similarly, Livy's *History*, books 1-10, is 159,132 words long but contains only 8,735 unique words, so its vocabulary density is 18.218.

Virgil's *Aeneid*, less than half as long (63,719 words), uses almost as many different words (7,531 of them), giving it a vocabulary density score of only 8.461. In other words, while Livy's vocabulary is larger than Virgil's, new words do not appear as frequently.

3 Things You Can Do with the Vocabulary Tool

The Perseus vocabulary tool is designed to be as versatile as possible, allowing both teachers and students to generate several kinds of vocabulary lists to help them teach or read Latin or Greek texts in the Perseus Digital Library. A few ways that this tool can be used are as follows:

- **A Comprehensive Vocabulary List for a Work**: An alphabetical list of all the words in a text serves as a text-specific dictionary, much like the glossary in the back of a student edition. Unlike the typical textbook glossary, however, this list also indicates which words are most common or most unique and therefore most deserving of a student's attention.
- **Pre-Reading or Orientation for a New Text or Author**: As noted above, the key term score identifies words that are relatively common in the texts in the vocabulary list but relatively rare in other documents in the digital library. From a pedagogical point of view, this list can be quite useful for orienting students to a new text or author. Because of the way that the key term score is defined, students are unlikely to have encountered the words with the highest key term scores often enough in their prior reading to have learned them. Thus, a list sorted by high key term scores showing at least the top fifty percent of words will be a very good guide for students to the unfamiliar words in a reading.
- **A List of Essential Words for an Author**: Advanced students working on mastery of a particular Greek or Latin author can make a list of all the words the author uses. A list of the top 40 or 50 percent, by weighted frequency, over all the author's works, will give students the essential words for reading that author. To list the author's most characteristic vocabulary, words relatively more common in his works than in other writers, select the same list of works, but sort by key term score, and look at the top 10%.

- **A List of Basic Words for Intermediate-Level Reading**: For an intermediate-level class, beginning to read unadapted texts, you can generate a word list based on the texts you are likely to assign first — Caesar and Cicero, Xenophon and Plato, or whatever your class will be reading. Select five or six texts or parts of texts, sort by weighted frequency, and request the top 50% or 60%. The list will be quite short, probably 200-500 words depending on the authors, and most of them will be familiar already, since the most frequent words in any given text are generally the most frequent words in the language as a whole. Once your students know all these words, they can begin reading, confident that they will know half to two-thirds of the words on a typical page. Many students are pleased and reassured to find out that the core vocabulary of ordinary Latin or Greek prose is so small.

- **A List of Essential Words for a Comprehensive Greek or Latin Exam**: Have your advanced students, who might be preparing for the Advanced Placement exam, comprehensive exams, or graduate-level qualifying exams, select the authors and works covered on the exam, requesting the top 70% or 80% by weighted frequency. This list will include all the most important vocabulary to know for the exam.

- **A List of Key Words for a Text**: If you want a quick overview of the potentially important words and concepts in a text, select the text that interests you with a sort order of key word score and a list size of top 10%. This will provide a short list of words for students to be aware of as they read the text.

- **On-line Vocabulary Review**: As noted at the outset, students learn vocabulary most effectively when they encounter words in context. Vocabulary lists displayed in an HTML table contain links to the Perseus word search tools. Thus a student wanting to study vocabulary can work through the list memorizing words while, at the same time, he or she can rapidly access the contexts in which the word appears.

4 Conclusions

Intermediate students often feel frustrated as they make the transition from their beginning textbook to actual Greek and Latin texts. They do not feel as though they are actually "read-

ing" the text because they are constantly looking words up in a dictionary. While focused vocabulary study, based on words they will actually encounter in their reading, is more efficient and more useful than rote memorization of abstract word lists, targeted vocabulary lists exist for only a few authors. The Perseus Vocabulary Tool addresses this problem by allowing students, teachers, and general readers to generate customized vocabulary lists for the texts they intend to read, allowing for more focused vocabulary study. More importantly, these lists can help students have a more enjoyable experience reading Greek and Latin texts because they will be less overwhelmed by unfamiliar vocabulary in their readings.

References

R. Harald Baayen. *Word Frequency Distributions*. Kluwer Academic Publishers, Dordrecht, 2001.

William E. Bull. Natural frequency and word counts: The fallacy of frequencies. *Classical Journal*, 44:469–484, 1948-1949.

James Coady. L2 vocabulary acquisition through extensive reading. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* Coady and Huckin [1997], pages 225–237.

James Coady and Thomas Huckin. *Second Language Vocabulary Acquisition: A Rationale for Pedagogy*. Cambridge, 1997.

Gregory Crane. Generating and parsing classical Greek. *Literary and Linguistic Computing*, 6(4):243–245, 1991.

Jan Hulstijn. Mnemonic methods in foreign language vocabulary learning: theoretical considerations and pedagogical implications. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* Coady and Huckin [1997], pages 203–224.

Kenneth F. Kitchell, Jr. Latin III's dirty little secret: Why Johnny can't read. *New England Classical Journal*, 24(4):206–226, 2000.

Batia Laufer. The lexical plight in second language reading: Words you don't know, words you think you know, and words you can't guess. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* Coady and Huckin [1997], pages 20–34.

Anne Mahoney. Studying the word study tool. *New England Classical Journal*, 28(3):181–183, 2001.

I. S. P. Nation. *Teaching and Learning Vocabulary*. Heinle and Heinle, Boston, 1990.

P. Nation and J. Coady. Vocabulary and reading. In R. Carter and M. McCarthy, editors, *Vocabulary and Language Teaching*, pages 97–110. Longman, London, 1988.

Kate Parry. Vocabulary and comprehension: two portraits. In *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* Coady and Huckin [1997], pages 55–68.

David J. Perry. Using the reading approach in secondary schools. In Richard A. LaFleur, editor, *Latin for the 21st Century: From Concept to Classroom*, pages 105–116. Scott Foresman Addison Wesley, Glenview, 1998.

Jeffrey A. Rydberg-Cox. Mining data from the electronic Greek lexicon. *Classical Journal*, Forthcoming 2001.

Gerard Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, Reading, MA, 1989.

Gerard Salton and Chris Buckley. Term-weighting approaches in automatic retrieval. *Information Processing and Management*, 1988.

Amit Singhal, Gerard Salton, and Chris Buckley. Length normalization in degraded text collections. *Proceedings of the Fifth Annual Symposium on Document Analysis and Information Retrieval*, 1996.

F. Tweedie and H. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Literary and Linguistic Computing*, 1998.

G. U. Yule. On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. *Biometrika*, 30:363–390, 1938.

G. U. Yule. *The Statistical Study of Literary Vocabulary*. Cambridge University Press, Cambridge, 1944.

George Kingsley Zipf. *The Psycho-Biology of Language: An Introduction to Dynamic Philology*. Houghton Mifflin Company, Boston, 1935.