Lexical and Semantic Relationships as Aids to Learnability: Cohesion and Reach

Leah GILNER
Franc MORALES

This paper is a report of results of two ongoing investigations that seek to contribute to the collection of measures employed to assess the adequacy of potential word lists for language instruction. In particular, these studies approach the topic from the point of view of the lexical and semantic relationships that the constituent words of word lists engage in. The constructs of *cohesion* and *reach* have been conceived and operationalized in order to measure and describe how words relate within a word list as well as with the rest of the lexicon. As the scope of this paper is limited to a report of results, it is relevant to provide an interpretative context. From the point of view of corpus analysis, please refer to Sinclair (1991, 1997), George (1997), Leech et al. (2001), Kilgarriff (1995), McEnery et al. (2006), Gilner and Morales (2008b), Biber (2006), and so on. From the point of view of the role of frequency in language use and learning, please refer to Nation (2004, 2006), Ellis (1996, 2001), Zahar (2001), Griffin (1998), Gilner and Morales (2008a), and so on. From the point of view word list design, please refer to West (1953), Faucett et al. (1936), Palmer (1931), Coxhead (2000), Richards (1974), Nation (1997, 2006), Bauer and Nation (1993), Laufer and Nation (1995), and so on. From the point of view of vocabulary acquisition, please refer to Carter (1998), Cobb et al. (2001), Horst et al. (1998), Hatch and Brown (1995), Schmitt (2000), Schmitt and Meara (1997), Wolter (2006), Stahl (1999), Stahl and Nagy (2006), Folse (2007), and so on.

The following two studies seek to characterize the lexical and semantic relationships in which General Service List (GSL; West, 1953) words and concepts participate. In particular, these studies investigate those relationships that exist among GSL words/concepts and those that exist between GSL words/concepts and the rest of the language. Thus, we speak of *cohesion* in reference to the former and *reach* in reference to the latter. These two concepts, cohesion and reach, are not single measures but a collection of measures: connectivity–density and activity–coverage. *Connectivity* considers whether or not two or more words/concepts are related while *density* considers in how many ways words/concepts are related. *Activity* considers only those words/concepts that participate in a particular relationship while *coverage* considers all words/concepts. In this manner, cohesion and reach are characterized by four distinct measures each. Note that this paper only reports on cohesion and reach in terms of connectivity and activity, that is, in terms of whether or not two words are related (rather than in how many ways they are related) and by considering only those words that participate in a particular relationship (rather than all words).

When we speak of the relationships between the GSL and the remainder of the language, we are referring to the language captured by a dictionary and, in particular, by a customized version of WordNet called ZETA. ZETA is a branch of version 2.0 of WordNet that has been modified extensively for the purposes of these studies. Unlike WordNet, for instance, ZETA contains entries for auxiliary verbs as well as the so-called 'function words', namely, prepositions, pronouns, determiners, and conjunctions. That said, these studies would not have been possible without the massive amount of information that ZETA has inherited from WordNet and, specifically, the extensive network of semantic and lexical relationships originates mostly in WordNet. Those familiar with WordNet will recognize, as approximates,

some of the figures used in these studies. Nonetheless, in order to avoid confusion, we will from now on speak only of ZETA and its characteristics although, as explained, many if not most of these have been derived from WordNet.

In its raw form, ZETA contains 144,430 lemmas referencing 203,636 senses (115,830 unique concepts). It also contains 1,053,458 relationships between words and 236,854 between concepts. For the purposes of these studies, ZETA was stripped of proper nouns. Thus if all senses of a lemma required capitalization (of the lemma), the lemma was removed. Similarly, if all referents of a concept required capitalization, the concept was removed. The remaining 106,008 lemmas and 97,127 concepts constitute the language database ($ZETA^{NP}$) from and against which we perform the analyses (705,767 relationships between words and 166,876 between concepts). Note that the GSL contains a proper noun – the word 'English' – which was removed from the list, leaving 2,283 words (rather than 2,284) for the following studies.

A lexical description of the GSL – along the parameters outlined – would be more informative if contrasted against other word lists. To this end, we fabricated 15 control lists out of two sources: the BNC corpus and ZETA. The BNC word lists were extracted from the raw unlemmatized frequency list provided by Leech et al. (2001). We first collapsed the list by part of speech and tallied frequency counts, eliminating 163,762 of the 794,771 entries in the list. The remaining 631,009 entries, however, included 188,430 proper nouns as well as 124,043 non-words (for instance, there were 53,482 entries that contained a semi-colon and 17,433 that contained a comma). These were eliminated. Finally, we removed those words that had no entries in ZETA, reducing the BNC word list to 55,543 words. Frequency information was kept throughout the process and allowed us,

at this point, to obtain two lists: a BNC^{FR} word list of 8,480 words (cut off at frequency 5 or more per million) and BNC^{IN} word list of 47,063 infrequent words.

The process of creation of the control word lists was identical for the lists originating in ZETA and BNC. Each word in ZETA was associated with a 12-digit random number and, then, the word-number pairs were sorted according to the random number, thus shuffling the words in an unpredictable manner. The first 2,283 words made the list ZETA-W_1. The entire operation was repeated twice more in order to obtain ZETA-W_2 and ZETA-W_3 and, in this manner, three random lists of words were obtained. The BNC^{FR} list and BNC^{IN} were likewise randomized and the first 2,283 words in each list yielded BNC-W^{FR} and BNC-W^{IN}, respectively.

Before describing the lexical relationships that the GSL words (and control lists) participate in, it is of interest to mention that the variation in the amount of polysemous words across lists and dictionary follows a pattern. Table 1 presents a breakdown of the data. Note that, for the sake of brevity, the results obtained from the control lists ZETA-W_1, ZETA-W_2, and ZETA-W_3 have been averaged in ZETA-W_{AVG}. This approach will be used throughout the remainder of the discussion.

	$ZETA^{NP}$		ZETA-W_{AVG}		GSL-W		BNC-W^{FR}		BNC-W^{IN}	
	mono	poly	mono	poly	mono	poly	mono	poly	mono	poly
Noun	63,438	12,973	1,349	275	332	1,362	481	1,085	1,026	388
Verb	6,102	5,195	137	110	236	1,032	185	589	135	105
Preposition	40	53	1	1	12	46	8	15	1	0
Pronoun	57	9	1	0	37	8	16	4	1	0
Adjective	15,411	5,289	345	116	133	439	249	409	490	201
Adverb	3,878	760	89	16	106	124	119	77	113	18
Determiner	25	3	1	0	17	3	9	1	0	0
Conjunction	28	14	0	0	20	13	10	4	0	0
Total	82,031	23,977	1,762	521	169	2,114	496	1,787	1,512	771

Table 1. Monosemy-Polysemy breakdown by part of speech.

The analysis showed that 22.62% of words in ZETA^{NP} are polysemous. The results obtained for ZETA-W_{AVG} (22.83%) were similar and somewhat less so for BNC-W^{IN} (33.78%). Striking, however, is the marked reversal in distribution shown by BNC-W^{FR} (78.28%) and especially by GSL-W (92.60%). Frequent words tend to have more senses than both infrequent words and the entire lexicon (Kilgarriff, 1997).

Also of interest, Table 2 presents a breakdown of each word set by part of speech. Note that a single word can have several parts of speech and that, therefore, the totals at the bottom of the table exceed the number of words in the lists and dictionary (unlike the previous analysis). The percentage columns show the amount of words (within each list) that have senses belonging to a particular part of speech and, therefore, cannot add up to 100%.

	ZETA^{NP}		ZETA-W_{AVG}		GSL-W		BNC-W^{FR}		BNC-W^{IN}	
	#	%	#	%	#	%	#	%	#	%
Noun	76,411	72.08%	1,624	71.13%	1,694	74.20%	1,566	68.59%	1,414	61.94%
Verb	11,297	10.66%	247	10.82%	1,268	55.54%	774	33.90%	240	10.51%
Preposition	93	0.09%	2	0.09%	58	2.54%	23	1.01%	1	0.04%
Pronoun	66	0.06%	1	0.04%	45	1.97%	20	0.88%	1	0.04%
Adjective	20,700	19.53%	461	20.19%	572	25.05%	658	28.82%	691	30.27%
Adverb	4,638	4.38%	105	4.60%	230	10.07%	196	8.59%	131	5.74%
Determiner	28	0.03%	1	0.04%	20	0.88%	10	0.44%	0	0.00%
Conjunction	42	0.04%	0	0.00%	33	1.45%	14	0.61%	0	0.00%
Total	113,275		2,441		3,920		3,261		2,478	

Table 2. Amount and percentage contribution of each part of speech.

The relative lack of polysemous words in ZETA^{NP}, ZETA-W_{AVG}, and BNC-W^{IN} is reflected by moderate increases in the size of each respective set. In contrast, GSL-W falls short of doubling its size while BNC-W^{FR} augments its size by almost half.

Looking at each part of speech, we can see that nouns are the most common, a relative frequency maintained across lists. Verbs, however, are significantly better represented among BNC-W^{FR} and especially among GSL-W, that is, lists made out of frequent words. More noticeable still, conjunctions, determiners, prepositions, and pronouns are up to 36 times more frequent in GSL-W than in the entire lexicon, as represented by $ZETA^{NP}$. Naturally, the actual amount of, for example, prepositions is larger in $ZETA^{NP}$ (n=93) than in GSL-W (n=58). It is the proportion of prepositions versus the entire set where the differences are appreciated.

Summing up, the distribution of GSL-W in terms of polysemy and part of speech appears to be influenced by the frequency of occurrence of its constituent words as shown in the correlations with BNC-W^{FR} as well as the inverse (polysemy) and lack of (part of speech) correlations with BNC-W^{IN}. Moreover, the distribution of GSL-W is atypical when one considers the lexicon at large ($ZETA^{NP}$) or random samples of it (ZETA-W_1, ZETA-W_2, and ZETA-W_3 averaged in ZETA-W_{AVG}).

As mentioned, we are only presenting here analyses of cohesion and reach in terms of connectivity and activity. Recall that connectivity considers whether or not two words participate in a relationship rather than in how many ways they do so, while activity considers only those words that participate in a relationship rather than all words. Thus, in terms of connectivity and activity, all values of cohesion and reach for $ZETA^{NP}$ are 100% and 0%, respectively, regardless of which relationship we consider. The lack of reach for $ZETA^{NP}$ should be evident. Words in $ZETA^{NP}$ are unable to form relationships with words outside of $ZETA^{NP}$ because there are no words outside of $ZETA^{NP}$. The full cohesion of $ZETA^{NP}$ in terms of activity should also be evident. Since activity only considers those words that engage in a particular relationship, all words that can form such a relationship are

thus considered and form the entire set under consideration. The contrast between activity and coverage provides further means to understand each concept. While, for example, 65.43% of nouns in ZETA^{NP} participate in relationships of synonymy (coverage), when taking into consideration only those words that participate in relationships of synonymy (activity), the cohesion value cannot be other than 100%. In this manner, neither cohesion nor reach for ZETA^{NP} in terms of connectivity-activity are informative and will not be included in the following analyses.

While the full study investigates each relationship separately (looking, for example, into the behavior of each part of speech per relationship), the summary offered here will be restricted to global measures. Thus, Table 3 presents the cohesion and reach values for all relationships in which GSL-W and the control lists participate as a function of the *share* of each in ZETA^{NP}. For instance, there are 72,068 words (out of 106,008) in ZETA^{NP} that participate in relationships of synonymy. Of these, 17.61% (share) are 'activated' by GSL-W words, 12.24% of which belong (cohe-

		ZETA-W$_{AVG}$		GSL-W		BNC-W^{FR}		BNC-W^{IN}	
		share	CO-RE	share	CO-RE	share	CO-RE	share	CO-RE
Synonymy	CO	6.12%	2.28%	17.61%	12.24%	13.63%	8.53%	6.38%	2.96%
	RE		97.72%		87.76%		91.47%		97.04%
Antonymy	CO	0.08%	22.22%	6.92%	88.91%	1.51%	73.00%	0.06%	50.00%
	RE		77.78%		11.09%		27.00%		50.00%
Hypernymy	CO	4.69%	5.14%	7.14%	30.45%	7.13%	21.45%	4.40%	3.79%
	RE		94.86%		69.55%		78.55%		96.21%
Hyponymy	CO	6.20%	3.90%	38.69%	5.62%	25.66%	5.96%	4.13%	4.04%
	RE		96.10%		94.38%		94.04%		95.96%
Holonymy	CO	3.23%	8.20%	10.87%	34.19%	7.04%	21.75%	2.56%	6.93%
	RE		91.80%		65.81%		78.25%		93.07%
Meronymy	CO	3.45%	7.79%	18.00%	20.64%	11.03%	13.89%	1.89%	9.42%
	RE		92.21%		79.36%		86.11%		90.58%

Table 3. Relationships between words.

sion) to GSL-W while 87.76% lay outside (reach) the GSL-W. In actual numbers, 1,553 GSL-W words are synonymous, that is, have at least one sense in common. Furthermore, 11,140 words outside of GSL-W participate in relations of synonymy with GSL-W words (in particular, with 2,053 GSL-W words).

The share allows us to determine the extent to which a list partakes in the total pool of words able to form relationships in ZETA^{NP}. The share is then proportionally divided among those words that contribute to the internal cohesion of a list and those words that are reached from a list. We can see, for example, that ZETA-W_{AVG} and BNC-W^{IN} share similar amounts of all possible synonyms in ZETA^{NP}, implying that a random collection of words offers similar activation as a collection of infrequent words. Conversely, frequent words have a larger share of all possible synonyms in the lexicon, roughly twice as large for BNC-W^{FR} and thrice for GSL-W. In actual numbers: GSL-W (re: 11,140; co: 1,553), ZETA-W_{AVG} (re: 4,313; co: 101), BNC-W^{IN} (re: 4,462; co: 136), and BNC-W^{FR} (re: 8,985; co: 838).

There are 6,643 words in ZETA^{NP} that participate in relationships of antonymy, 83,352 words that participate in relationships of hypernymy-hyponymy, and 11,184 words that participate in relations of holonymy-meronymy. The share values (of these amounts) by word list again indicate that infrequent and random words have a lower capacity to engage other words. Briefly, two observations are of interest. First, the share of antonyms in GSL-W is not only significantly greater than that of other lists but it is also strongly biased towards cohesion, that is, antonym relationships among GSL-W words. An implication of this is that GSL-W words appear to be more readily able to convey opposition. Second, comparing the values for hypernym-hyponym relationships across lists, it is possible to assert that frequent words tend to be more general than infrequent or random words,

since they can engage a much larger share of hyponyms (more specific words) than hypernyms (more general words).

Table 4 shows the total share of lists across lexical relationships by part of speech. The last row, labeled 'Global', indicates the corresponding shares of all words in ZETA^{NP} that participate in any of the lexical relationships mentioned previously. Thus, a list of random words (ZETA-W_{AVG}) engages 12.72% of all possible words while a list of infrequent words (BNC-W^{IN}) engages 11.16%.

		ZETA-W_{AVG}		GSL-W		BNC-W^{FR}		BNC-W^{IN}	
		share	CO-RE	share	CO-RE	share	CO-RE	share	CO-RE
Noun	CO	12.71%	2.37%	43.83%	4.57%	31.55%	4.70%	10.82%	2.42%
	RE		97.63%		95.43%		95.30%		97.58%
Verb	CO	21.86%	2.49%	88.85%	12.35%	64.11%	8.47%	15.69%	1.86%
	RE		97.51%		87.65%		91.53%		98.14%
Preposition	CO	5.08%	0.00%	81.36%	70.83%	52.54%	19.35%	1.69%	0.00%
	RE		100.00%		29.17%		80.65%		100.00%
Pronoun	CO	0.00%	0.00%	91.67%	81.82%	75.00%	22.22%	0.00%	0.00%
	RE		0.00%		18.18%		77.78%		0.00%
Adjective	CO	5.86%	2.11%	14.19%	15.17%	12.17%	10.95%	8.84%	3.49%
	RE		97.89%		84.83%		89.05%		96.51%
Adverb	CO	4.24%	2.93%	15.72%	23.52%	13.89%	12.53%	4.07%	1.53%
	RE		97.07%		76.48%		87.47%		98.47%
Determiner	CO	4.76%	0.00%	85.71%	83.33%	57.14%	100.00%	0.00%	0.00%
	RE		100.00%		16.67%		0.00%		0.00%
Conjunction	CO	0.00%	0.00%	93.33%	78.57%	40.00%	50.00%	0.00%	0.00%
	RE		0.00%		21.43%		50.00%		0.00%
Global	CO	12.72%	2.38%	42.62%	4.89%	31.36%	4.98%	11.16%	2.51%
	RE		97.62%		95.11%		95.02%		97.49%

Table 4. All lexical relationships by part of speech.

In actual numbers, there are 100,584 words out of 106,008 in ZETA^{NP} that participate in at least one of the lexical relationships discussed. Of these, GSL-W engages 40,776 or 42.62% while BNC-W^{FR} engages 29,792

or 31.36%. Speaking in terms of cohesion, 2,097 out of 2,283 (91.85%) words in GSL-W are engaged in internal lexical relationships, compared with 1,571 out of 2,283 words (68.81%) for BNC-W^{FR}, 282 out of 2,283 words (12.35%) for BNC-W^{IN}, and 305 out of 2,283 (13.35%) words for ZETA-W_{AVG}. Speaking in terms of reach, GSL-W and control lists can potentially engage the same number of words, namely, 98,301. However, GSL-W reaches 40,776 words, while BNC-W^{FR} reaches 29,972 words, BNC-W^{IN} reaches 10,939 words, and ZETA-W_{AVG} reaches 12,488 words.

The breakdown by part of speech shows that GSL-W is comparatively more active than any other list, also demonstrating higher cohesion and reach. Furthermore, function words and verbs are manifestly well connected in GSL-W. This is also true in BNC-W^{FR} but to a lesser extent. Last, and as expected from previous results, lists of random words and infrequent words display the least amount of cohesion and capacity to reach out to the remainder of the lexicon.

We now turn our attention to the study that seeks to characterize the semantic relationships in which GSL concepts engage. As mentioned $ZETA^{NP}$ contains 97,127 concepts and exhibits 166,876 relationships between these concepts. An inspection of the GSL reveals that it contains 14,230 concepts. The following discussion will explore in which manner these concepts relate among themselves, on the one hand, and with the remaining concepts in the language, on the other.

In the lexical analysis, control lists contained the same number of words as GSL-W. The control lists manufactured for this study contain the same number of concepts as GSL-C. Beyond that, the process of clustering and randomization was identical. Thus, BNC-C^{FR} and BNC-C^{IN} represent lists of concepts referred to by frequent and infrequent words, respectively, while ZETA-C_{AVG} is the average of the random list of concepts ZETA-C_1,

ZETA-C_2, and ZETA-C_3 extracted from the concept database in ZETA^{NP}. A minimal selection of analyses from the full study follows.

Table 5 shows a breakdown of the noun concepts under investigation according to a number of semantic categories. These categories have been sorted from more to less abundant in ZETA^{NP}. Two columns are given for each set. The left column indicates the actual number of noun concepts in each category while the right column shows the proportional contribution

	ZETA^{NP}		ZETA-C_{AVG}		GSL-C		BNC-C^{FR}		BNC-C^{IN}	
Artifact	10,803	17.50%	1,572	17.35%	1,109	17.91%	1,173	16.99%	1,429	17.80%
Act	6,136	9.94%	926	10.22%	826	13.34%	1,009	14.61%	913	11.37%
Person	5,776	9.36%	857	9.46%	421	6.80%	618	8.95%	1,071	13.34%
Plant	5,134	8.32%	735	8.11%	74	1.20%	88	1.27%	361	4.50%
Communication	4,312	6.99%	623	6.88%	628	10.14%	732	10.60%	530	6.60%
Animal	4,097	6.64%	601	6.63%	93	1.50%	123	1.78%	468	5.83%
State	3,367	5.45%	506	5.59%	302	4.88%	329	4.77%	463	5.77%
Attribute	2,841	4.60%	423	4.67%	399	6.44%	408	5.91%	534	6.65%
Substance	2,802	4.54%	398	4.40%	120	1.94%	150	2.17%	365	4.55%
Cognition	2,584	4.19%	373	4.12%	370	5.98%	385	5.58%	314	3.91%
Food	2,293	3.71%	334	3.68%	113	1.83%	143	2.07%	273	3.40%
Body	1,926	3.12%	277	3.05%	117	1.89%	125	1.81%	178	2.22%
Group	1,768	2.86%	268	2.95%	298	4.81%	309	4.48%	139	1.73%
Quantity	1,023	1.66%	151	1.67%	159	2.57%	156	2.26%	146	1.82%
Possession	1,019	1.65%	145	1.60%	130	2.10%	138	2.00%	71	0.88%
Event	1,012	1.64%	146	1.62%	218	3.52%	240	3.48%	155	1.93%
Location	785	1.27%	127	1.40%	171	2.76%	143	2.07%	76	0.95%
Object	762	1.23%	113	1.24%	142	2.29%	121	1.75%	116	1.44%
Process	738	1.20%	107	1.18%	41	0.66%	82	1.19%	118	1.47%
Time	708	1.15%	106	1.17%	155	2.50%	119	1.72%	79	0.98%
Phenomenon	614	0.99%	95	1.05%	69	1.11%	65	0.94%	70	0.87%
Feeling	408	0.66%	60	0.66%	82	1.32%	80	1.16%	72	0.90%
Relation	401	0.65%	62	0.68%	55	0.89%	73	1.06%	41	0.51%
Shape	330	0.53%	46	0.50%	62	1.00%	71	1.03%	45	0.56%
Tops	45	0.07%	4	0.05%	27	0.44%	18	0.26%	2	0.02%
Motive	41	0.07%	7	0.07%	10	0.16%	6	0.09%	1	0.01%
Total	61,725	100.00%	9,060	100.00%	6,191	100.00%	6,904	100.00%	8,030	100.00%

Table 5. Breakdown of noun concepts by semantic category.

each category makes to its corresponding set.

The row labeled ‘Total’ presents the total number of concepts available in ZETA^{NP} as well as in each of the concept lists. It is noteworthy that both GSL-C and BNC-C^{FR} have significant lower amounts of noun concepts when compared against BNC-C^{IN} and ZETA-C$_{AVG}$, that is, concept lists of equal size. As will be shown later on, part of the reason may lie in the fact that verb concepts are comparatively much more frequent in GSL-C and, to a lesser extent, in BNC-C^{FR}.

A quick glance at Table 5 reveals that, in general, similar trends are manifested by all word lists in accordance with noun concept distributions in ZETA^{NP}. Differences, however, exist. The category ‘Plant’ is comparatively underrepresented in GSL-C and BNC-C^{FR}, as is the category ‘Animal’ and ‘Person’, while the categories ‘Communication’, ‘Attribute’, ‘Group’, and ‘Event’ are comparatively overrepresented. The category ‘Tops’ refers to noun concepts that stand as beginner primitives in the hypernym-hyponym relationship, that is, as the most general concepts from which hierarchies of gradually more specific noun concepts in ZETA^{NP} derive. Of particular relevance, GSL-C accounts for over half (60%) of the noun concepts available in this category in ZETA^{NP} and BNC-C^{FR} for 40%. Numbers drop to 0.04% in BNC-C^{IN} and 0.09% in ZETA-C$_{AVG}$ despite the fact that these two concept lists are, over all, comparatively larger.

Table 6 presents a breakdown of verb concepts by semantic category. As before, the row labeled ‘Total’ presents the total verb concepts per set and shows that GSL-C contains 37.04% of all verb concepts despite the entire concept list accounting for only 14.65% of ZETA^{NP}. Verb concepts are unusually numerous in GSL-C.

As is the case with noun concepts, all breakdowns display similar trends of distribution. The concept list ZETA-C$_{AVG}$ shows the closest proportional

	ZETA^{NP}		ZETA-C$_{AVG}$		GSL-C		BNC-C^{FR}		BNC-C^{IN}	
Change	2,313	17.10%	351	17.75%	704	14.05%	579	14.64%	336	19.82%
Contact	2,158	15.96%	305	15.44%	748	14.93%	641	16.21%	284	16.76%
Communication	1,533	11.33%	215	10.89%	562	11.22%	465	11.76%	189	11.15%
Motion	1,374	10.16%	218	11.04%	494	9.86%	410	10.37%	152	8.97%
Social	1,095	8.10%	158	7.98%	433	8.64%	324	8.19%	143	8.44%
Possession	806	5.96%	114	5.75%	316	6.31%	235	5.94%	78	4.60%
Stative	750	5.55%	111	5.60%	391	7.80%	269	6.80%	69	4.07%
Cognition	688	5.09%	102	5.16%	305	6.09%	241	6.10%	82	4.84%
Creation	674	4.98%	95	4.81%	228	4.55%	170	4.30%	87	5.13%
Body	537	3.97%	83	4.22%	177	3.53%	143	3.62%	85	5.01%
Competition	456	3.37%	64	3.22%	172	3.43%	138	3.49%	38	2.24%
Perception	451	3.33%	60	3.03%	178	3.55%	119	3.01%	62	3.66%
Emotion	337	2.49%	48	2.43%	144	2.87%	92	2.33%	51	3.01%
Consumption	243	1.80%	36	1.80%	104	2.08%	87	2.20%	30	1.77%
Weather	82	0.61%	14	0.71%	30	0.60%	29	0.73%	8	0.47%
Auxiliary	28	0.21%	3	0.17%	24	0.48%	12	0.30%	1	0.06%
Total	13,525	100.00%	1,977	100.00%	5,010	100.00%	3,954	100.00%	1,695	100.00%

Table 6. Breakdown of verb concepts by semantic category.

agreement with ZETA^{NP}. GSL-C and BNC-C^{FR} again display the largest divergences. The categories ‘Change’ and ‘Contact’ are less sloped among the frequent lists, disseminating concepts over categories less favored by ZETA^{NP} as well as the infrequent and random lists. In particular, the categories ‘Stative’ and ‘Cognition’ receive the largest endorsements.

Table 7 presents share, cohesion, and reach values exhibited by the list of concepts across the semantic relationships under consideration (naturally, the relationship of synonymy does not apply to concepts). In regards to antonymy, percentages of share across concept lists partake in similar amounts from the pool of 7,424 available concepts in ZETA^{NP} that participate in this kind of relationship. Interestingly, cohesion and reach values are opposed for GSL-C and BNC-C^{FR} versus BNC-C^{IN} and ZETA-C$_{AVG}$.

In regards to hypernymy and hyponymy, differentials across concept lists are in agreement with those observed earlier across word lists (refer to Table

		ZETA-C_{AVG}		GSL-C		BNC-C^{FR}		BNC-C^{IN}	
		share	CO-RE	share	CO-RE	share	CO-RE	share	CO-RE
Antonymy	CO	15.20%	15.01%	19.64%	57.89%	19.38%	30.85%	19.05%	19.66%
	RE		84.99%		42.11%		69.15%		80.34%
Hypernymy	CO	10.67%	30.18%	13.34%	76.04%	11.74%	63.29%	9.05%	27.14%
	RE		69.82%		23.96%		36.71%		72.86%
Hyponymy	CO	15.80%	20.41%	43.15%	23.51%	36.12%	20.58%	11.69%	21.00%
	RE		79.59%		76.49%		79.42%		79.00%
Holonymy	CO	11.08%	26.84%	13.73%	54.06%	12.36%	43.41%	9.75%	23.52%
	RE		73.16%		45.94%		56.59%		76.48%
Meronymy	CO	12.04%	24.68%	21.05%	35.26%	17.94%	29.91%	9.59%	23.92%
	RE		75.32%		64.74%		70.09%		76.08%

Table 7. Relationships between concepts.

3). Frequent lists have more hyponyms than infrequent and random lists. In frequent lists, the majority of hypernyms belongs to internal relationships (cohesion) while in infrequent and random lists, the majority belongs to external relationships (reach). In other words, GSL-C and BNC-C^{FR} are composed of concepts that, over all, are more general than of those in BNC-C^{IN} and ZETA-C_{AVG}.

As shown throughout the presentation of results from these two studies, the constructs of *cohesion* and *reach* are informative as well as adequate for the assessment of potential word lists in the realm of lexical and semantic relationships. As is always the case when speaking about selection, the rational design of word lists is of fundamental importance. Nation and Macalister (2007) make a strong case for the necessity to adhere to coherent and founded principles of selection, positing that most instruction fails precisely because of shortcomings in this area.

References

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Biber, D. (2006). *Longman grammar of spoken and written English*. Harlow: Longman.

Carter, R. (1998). *Vocabulary: applied linguistic perspectives*. New York: Routledge.

Cobb, T., Greaves, C., & Horst, M. (2001). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la didactique des langues secondes* (pp. 133–153). Montréal: Éditions logique.

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, *34*(2), 213–238.

Ellis, N. (1996). Sequencing in SLA: Phonological Memory, Chunking, and Points of Order. *Studies in Second Language Acquisition*, *18*, 91–126.

Ellis, N. (2001). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Faucett, L., Palmer, H., Thorndike, E. L., & West, M. (1936). *Interim Report on Vocabulary Selection*. London: P. S. King and Son, Ltd.

Folse, K. S. (2007). *Vocabulary Myths: Applying second language research to classroom teaching*. Ann Arbor: The University of Michigan Press.

George, H. V. (1997). *Essays in Informational English Grammar with reference to English language teaching*. Victoria, AU: La Trobe University.

Gilner, L., & Morales, F. (2008a). Elicitation and application of a phonetic description of the General Service List. *System*, *36*(4). doi:10.1016/j.system.2008.02.006

Gilner, L., & Morales, F. (2008b). Corpus-based frequency profiling: Migration to a word list based on the British National Corpus. *The Buckingham Journal of Language and Linguisitics*, *1*, 41–58.

Griffin, Z. M., & Bock, K. (1998). Constraint, Word Frequency, and the Relationship between Lexical Processing Levels in Spoken Word Production. *Journal of Memory and Language*, *38*, 313–338.

Hatch, E. M., & Brown, C. (1995). *Vocabulary, semantics, and language education*. New York: Cambridge University Press.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond A Clockwork Orange: Acquiring second-language vocabulary through reading. *Reading in a Foreign Language*,

11, 207–223.

Kilgarriff, A. (1995). BNC database and word frequency lists. from http://www.kilgarriff.co.uk/bnc-readme.html

Kilgarriff, A. (1997). Putting Frequencies in the Dictionary. I*nternational Journal of Lexicography*, *10*, 135–155.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*, 307–322.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English : based on the British National Corpus*. Harlow: Longman.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. London; New York: Routledge.

Nation, I. S. P. (2004). study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language* (pp. 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How Large a Vocabulary is Needed for Reading and Listening? *Canadian Modern Language Review*, *63*(1), 59–82.

Nation, I. S. P., & Macalister, J. (2007). *Language Curriculum Design* (2nd ed.). Wellington: School of Linguistics and Applied Language Studies, Victoria University.

Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: description, acquisition and pedagogy* (pp. 6–19). Cambridge: Cambridge University Press.

Palmer, H. (1931). *Second Interim Report on Vocabulary Selection submitted to the Eighth Annual Conference of English Teachers under the auspices of the Institute for Research in English Teaching*. Tokyo: IRET.

Richards, J. C. (1974). Word lists: problems and prospects. *RELC Journal*, *5*(2), 69–84.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge; New York: Cambridge University Press.

Schmitt, N., & Meara, P. (1997). Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in Second Language Acquisition*, *19*, 17–36.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University

Press.
Sinclair, J. (1997). *Collins Cobuild English grammar*. London: Harper Collins.
Stahl, S. A. (1999). *Vocabulary development*. Cambridge, Mass.: Brookline Books.
Stahl, S. A., & Nagy, W. E. (2006). *Teaching word meanings*. Mahwah, N.J.: L. Erlbaum Associates.
West, M. (1953). *A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London; New York: Longmans, Green.
Wolter, B. (2006). Lexical network structures and L2 vocabulary acquisition: The role of L1 lexical/conceptual knowledge. *Applied Linguistics*, *27*(4), 741–747.
Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, *57*(4), 541–572.