Chapter 2

Is there room for an academic word list in French?

Tom Cobb and Marlise Horst
Université du Québec à Montréal, Concordia University

Abstract

Extensive analysis of corpora has offered learners of English a solution to the problem of which among the many thousands of English words are most useful to know by identifying lists of high frequency words that make up the core of the language. Of particular interest to university-bound learners is Coxhead's (2000) Academic Word List (AWL). Analyses indicate that knowing the 570 word families on this list along with the 2000 most frequent families consistently offers coverage of about 85% of the words learners will encounter in reading an academic text in English. This finding raises the question of whether such lists can be identified in other languages. The research reported in this chapter provides an initial answer in the case of French. Lists of the 2000 most frequent French word families were built into an online lexical frequency profiling program (*Vocabprofil*) and their coverage powers tested. Analyses of texts using this tool confirmed the usefulness of the lists in identifying distinct and consistent profiles for French texts of three specific genres (newspaper, popular expository, and medical). Comparisons using parallel French and English texts indicated that the 2000 most frequent word families of French offer the reader a surprisingly high level of coverage (roughly 85%), a level that can only be achieved in English with the knowledge of the most frequent 2000 words plus the 570 AWL words. In other words, the French 2000 list seems to serve both everyday and academic purposes more effectively than its English counterpart, such that there appears to be no need for an additional AWL-like list in French to facilitate the comprehension of academic texts. With the coverage powers of the French 2000 list so high, there appears to be little or no space left in the lexis of French for such a list to occupy.

1. Introduction

Acquiring a second lexicon is a daunting task for language learners, especially if the goal is to achieve literacy in the second language. But the task becomes more manageable if we know which words are more important to learn than others, or which words are most useful to know as a precondition to learning others. In English, computational studies of word frequency and text coverage, in conjunction with empirical studies of learner comprehension of texts with different lexical profiles, have provided valuable information for both course designers and independent learners. It has become clear that words of particular frequencies have predictable degrees of prominence in texts of particular genres. For example, the 1000 most frequent words, along with proper nouns, tend through repetition to make up, or cover, about 90% of the running words in spoken conversations. This type of analysis, known as lexical frequency profiling (or LFP, Laufer & Nation 1995), has been useful in clarifying and resolving specific problems of lexical acquisition in English. Particularly interesting is the Academic Word List (or AWL, Coxhead 2000) component of the LFP framework, which combines frequency, coverage, and genre information to provide learners with a useful solution to a problem in the naturalistic acquisition of the vocabulary needed for reading academic texts.

An interesting question, then, is whether LFP analysis is applicable to languages other than English. While frequency lists have been developed over the years for most European languages, neither the coverage properties nor the genre determination of these lists have been closely examined. This chapter is a preliminary comparison of the vocabulary distributions of English and French using the LFP framework, with emphasis on the question of whether or not there is any lexical zone in French resembling the English AWL that might be useful to learners of French. We begin our investigation with some background on the nature of the problem the AWL resolves for learners of English.

2. A logical problem in acquiring some second lexicons

Whether the number of words in a modern language is 50,000 or 500,000 or somewhere in between, as variously claimed by different researchers using different counting units and methods, either of these numbers is daunting to an L2 learner. There are at least three factors working against the acquisition of a second lexicon in English. First, many learners are simply unlikely ever to

meet a large proportion of the lexicon. The vast majority of English words are found mainly in written texts, while a relatively small handful are encountered in daily conversation and watching television. This means that for the many learners who achieve conversational fluency in an L2 rather than full literacy, the vast majority of words are simply inaccessible for learning through naturalistic acquisition.

Second, even for avid readers, vocabulary acquisition through exposure to texts is slow and uncertain. The classic finding is that there is only .07% likelihood of a first language learner later recognizing the meaning of a new word after encountering it once incidentally in reading (Nagy & Herman 1987). This rate is nonetheless adequate to explain the attainment of an adult-sized English lexicon (defined as about 20,000 word families) based on an average reading program of 1 million words a year. But few L2 learners are likely to read this much; the highest estimate we know of for an extensive L2 reading program is 300,000 words per year, (personal communication from R. Rozell, teaching in Japan 2002).

Third, the probability of word learning from reading is likely to be even lower than .07% for L2 readers. Natural acquisition relies on new words being met in environments where most of the surrounding words are known, as will normally be the case for school-age learners meeting new words in level-appropriate texts in their own language. For L2 readers, however, unknown words are likely to arrive not alone but in clusters. A typical inference exercise in the ESL classroom is to work out the meaning of *date* from lexically dense sentences like, 'Her date eventually motored into view well past the ETA'. The problem of learning from such contexts is real and quantifiable. Research has shown that the minimal ratio of known to unknown words for both reliable comprehension and new acquisition is at least 20 : 1, or in other words when at least 95% of the running words in the environment are known (see Nation 2001 for an overview). These circumstances are unlikely to prevail when L2 learners read unsimplified texts. Thus, the natural acquisition of English as a second lexicon presents a problem: the numbers simply do not add up.

Let us look in detail at a typical learner's progress toward knowing 95% of the running words in an average text. Whether in a classroom or a naturalistic setting, learners tend to acquire L2 vocabulary in rough order of frequency. The 1000 most frequent word families (base words along with their most common derivations and inflections) of English are very frequent in spoken language, and in addition account for around 75% of the running words in most kinds of written language[1] so opportunities for meeting and learning

these words are good. Then, learners who read or who join reading-based courses are likely to acquire some or most of the second thousand most frequent words. Words in this category are relatively infrequent in speech but occur often in writing, accounting for another 5% of the running words in many text types. So with just 2000 known word families, the learner already controls about 80% of the running words in an average text. If this rate of return could be sustained (roughly 5% additional coverage per additional 1000 word families learned), then the trajectory from 80 to 95% coverage would be achieved with knowledge of 5000 English word families. But in fact, coverage does not proceed linearly in neat 5% increases with each 1000 words learned. Unfortunately, laying the vocabulary basis for naturalistic acquisition of less frequent items is not so easily accomplished, at least not in the case of English.

It turns out that in natural texts, the chances of meeting less common words drop off rather sharply after the 2000 word zone. Table 1 shows the typical coverage percentages provided by the different frequency bands — percentages for the ten most frequent words (*the, a, of, I)* appear at the bottom of the table, with those for the 100 most frequent words (*house, big, way, girl*) just above, and so on. As mentioned, the 1000 most frequent words are seen to cover about 75% of the words in an average text, and 2000 cover just over 80%.

The coverage curve rises steeply, levels off at around 80%, and thereafter creeps only very gradually towards the 95% mark, as is clearly evident when this information is represented graphically. As Figure 1 shows, the fourth thousand most frequent words account for just an additional 3% of running words, the fifth an additional 1%, and so on — with the 95% mark receding into the distance at 12,000 words. A similar too-much-to-learn problem probably

Table 1. Typical coverage figures for different frequency bands (Carroll, Davies and Richman 1971, cited in Nation 2001).

Number of words	Text coverage
87,000	100%
44,000	99%
12,000	95%
5,000	89%
4,000	88%
3,000	85%
2,000	81%
1,000	74%
100	49%
10	24%

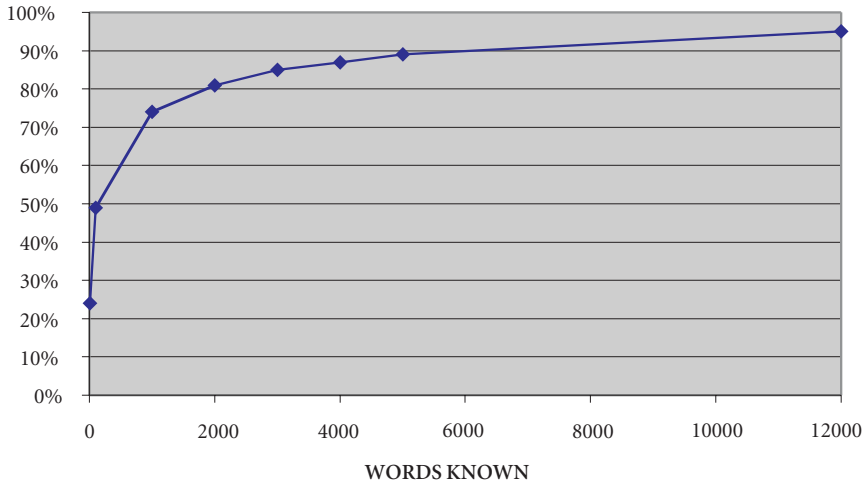

Figure 1. Graphic representation of part of Table 1.

features in any language where a significant proportion of the lexicon is housed mainly on paper. Hazenberg & Hulstijn (1996) reached this conclusion in the case of Dutch, finding that the minimal lexical base for reading demanding texts was knowing 90% of their running words, and that achieving this meant knowing roughly 10,000 word families. For most learners of either language, the 95% or even 90% coverage mark will not be achieved by simple cumulative progression through the frequency levels via courses, contextual inference, dictionary look-ups, or any of the other usual means of natural vocabulary growth. The number of words to learn is simply overwhelming.

To summarize, the deck is stacked against a L2 learner acquiring a functional reading lexicon in English or possibly in any language. It seems the time required for natural acquisition of a second lexicon is approximately all the time available, that is, the time it takes to grow up and be educated in a language. This is time that most L2 learners simply do not have. The natural frequency distribution of words will incline many L2 learners to plateau when they have acquired the resources needed for basic spoken interaction; those who attempt to proceed toward fuller literacy will find the going hard because of the number of words needed for fluent reading and the inhospitable conditions of acquiring them. And yet a large proportion of the world's English learners, if not the majority, are studying English precisely in order to read documents in their academic, professional, or vocational areas.

It is often said there is a 'logical problem' with language acquisition, meaning there is more to be learned than there are learning resources available (Baker & McCarthy 1981; Gold 1967; Pinker 1995). This analysis has been applied to the acquisition of syntax, but here we see that some version of it applies to the acquisition of the lexicon as well, at least the English lexicon for an L2 learner wishing to read fluently. The quantity to be learned cannot be accounted for in terms of known models of naturalistic acquisition. Unfortunately, the innatist solution proposed for syntax has not offered much with regard to lexis on either the explanatory or practical level, so more mundane solutions have been sought.

3. The AWL as a solution to the problem

A partial solution to the problem of building a second lexicon in English has been found in the form of the AWL (Coxhead 2000). This is a list of 570 words such as *abandonment, abstraction,* and *accessible* that, while not necessarily frequent in the language at large, have been found through extensive corpus analysis to be frequent across the genre of academic texts (Xue & Nation 1984; Coxhead 2000). The AWL, along with the 1000 and 2000 frequency lists, form a combined word list of 2570 families that gives reliable coverage of about 90% of the running words in an academic or quality newspaper text. In other words, the AWL adds another roughly 10% coverage for an additional learning investment of only 570 word families. Learning the meanings of the 570 AWL items presents a challenging but feasible task, and several instructional ideas have been proposed for helping ESL learners get control of its contents (e.g., activities on Coxhead's *Academic Word List* and Cobb's *Lexical Tutor* websites).

The substantial additional coverage that the AWL offers is clear but the reader will recall that the critical point for independent reading and further acquisition is not 90 but 95% coverage. Nation and his colleagues have argued that further coverage shortcuts can be discovered in the relatively definable and manageable lexicons that exist within particular domains, for example in economics (Sutarsyah, Nation & Kennedy 1994). Thus the additional word learning that takes a learner to the 95% criterion will occur more or less naturalistically through content instruction within such domains. The role of the AWL is to provide a reliable bridge between these two relatively accessible zones of lexical acquisition, between the words that are frequent in the language at large and the words that are frequent within a specific domain of study.

The reliability of the AWL's coverage can be demonstrated with the help of a computer program that analyzes texts in terms of their lexical frequency profiles. This analysis is a key component of the LFP framework and is useful both for validating coverage claims and exploiting coverage information instructionally (e.g., for assessing lexical richness in learner productions, and determining the lexical density of reading materials). To analyze a text, the program simply loads the 1000, 2000, and AWL family lists into memory and then classifies each word according to the frequency list it belongs to. Words not on any list are designated off-list. The program keeps tallies of the number of items in each category and calculates the proportion each category represents. A typical output for a newspaper text might show that 70% of the total number of running words are among the 1000 most frequent words, 10% are from the 1001–2000 list, 10% are from the AWL, leaving a rump of 10% off-list items that include a mix of proper nouns and lower frequency items. Various versions of this computer program exist, but for purposes of public replicability the first author's Internet-based version of the program on the *Lexical Tutor* website, *Web Vocabprofile* (henceforth referred to as VP), will be used in the analyses to follow. These analyses will serve two purposes, first to show what the English AWL is and does, and second to provide a specific baseline to measure the success of exporting the LFP framework to a language other than English. The second part of this chapter reports our initial foray into developing and testing the framework for French.

To demonstrate the AWL's reliability and coverage, we submitted a set of seven 2000+ word text segments from the *Learned* section of the Brown corpus (Francis & Kucera 1979) to VP analysis. The results of these analyses are shown in Table 2. While the texts represent a range of disciplines, coverage percentages are remarkably similar. Mean AWL coverage is 11.60%, which as predicted, combines with the coverage provided by the 1000 and 2000 lists to amount to a reasonably reliable 90% figure. Reliability of coverage is evident in standard deviations from means for all four frequency categories: these are small, on the order of 2 to 4%. *Chi*-square comparisons show no significant differences across disciplines, with the exception of medicine-anatomy (possibly owing to the very high proportion of specialist terminology in this field).

A second property of the AWL that can be demonstrated with VP is its genre determination. To make a simple cross-genre comparison, we chose a second genre, that of popular expository writing designed for the general reader. We predicted that texts of this genre would also provide consistent lexical profiles but with smaller contributions from the AWL than was found

Table 2. Lexical frequency profiles across disciplines (coverage percentages).

Brown segment	Discipline	No. of words	1000	2000	1000 + 2000	AWL	1K + 2K + AWL
J32	Linguistics	2031	73.51	8.37	81.88	12.60	94.48
J29	Sociology	2084	74.23	4.75	78.98	13.44	92.42
J26	History	2036	69.3	5.7	75.00	14.49	89.49
J25	Social Psychology	2059	73.63	3.11	76.74	14.38	91.12
J22	Development	2023	76.42	4.55	80.97	12.26	93.23
J12	Medicine (anatomy)	2024	71.05	3.80	74.85	6.72	81.57
J11	Zoology	2026	75.12	6.17	81.29	7.31	88.60
M			73.32	5.21	78.53	11.60	90.13
SD			2.42	1.74	3.01	3.24	4.30

Note 1: In this and subsequent Tables 1K and 2K refer to the first and second thousand word lists respectively
Note 2: Segments from the Brown corpus are described in the Brown University website accessible from the AWL page on the *Lexical Tutor* website.

for the academic texts from the Brown corpus. We profiled 17 randomly selected non-fiction *Reader's Digest* articles of around 200 words each. The results in Table 3 show that coverage percentages are similar across texts on widely differing topics, and that once again standard deviations for category means are low. As expected, the role of AWL items is less prominent (5.56% or about half of what it was in the academic texts). The significant difference found in a *chi*-square comparison for the means of the popular and scientific texts shows that the two genres have distinct lexical profiles.

Other dimensions of the AWL that can be investigated with the help of VP include hypotheses about the list's semantic content. It is arguable that AWL items are not only important because they increase text coverage, but also because of the intellectual work they do in academic texts. These words appear frequently across academic domains for the good reason that they are used to *define*, *delineate*, *advance*, and *assess abstract entities* such as *theories*, *arguments*, and *hypotheses* (the italicized words are AWL items). Researchers in the field of L1 literacy such as Olsen (1992) and Corson (1997) stress the need for English-speaking children to learn to use these Greco-Latin words and to think with the concepts they represent; they argue that a 'lexical bar' faces those who fail to do so (Corson 1985). VP analysis was used in the following way to test the claim that the language of theories and arguments is largely AWL language: the second author performed a *Gedankenexperiment* by typing into the *Vocabprofile* Text

Table 3. Consistent but distinct profiles for non-academic texts (coverage percentages).

TEXT	1000	2000	1000+ 2000	AWL	1K + 2K + AWL
Audubon	71.62	9.46	81.08	4.73	85.81
Computers	76.62	4.98	81.60	10.95	92.55
Drugs	74.46	5.98	80.44	3.26	83.70
Earthquakes	70.89	10.13	81.02	7.59	88.61
Dieting	75.62	6.97	82.59	3.98	86.57
Gas	89.47	0.00	89.47	7.02	96.49
Olympics	72.55	3.92	76.47	5.88	82.35
Origins of Life	69.71	7.43	77.14	4.00	81.14
Plague	75.76	3.03	78.79	5.45	84.24
Salt	81.60	5.66	87.26	3.30	90.56
Stage fright	76.44	9.13	85.57	4.33	89.90
Teenagers	84.03	4.86	88.89	4.17	93.06
Tennis	70.00	7.89	77.89	5.26	83.15
Toledo	69.05	4.76	73.81	6.55	80.36
Vitamins	67.30	9.43	76.73	6.29	83.02
Volcanoes	75.63	7.11	82.74	4.06	86.80
Warm-Up	73.33	6.67	80.00	7.78	87.78
M	74.95	6.32	81.26	5.56	86.83
SD	5.73	2.62	4.47	1.99	4.56

Entry box all the discourse or argument structuring words that came to mind in five minutes and then submitting these for analysis. The items produced by this procedure were as follows:

> concede imply hypothesize infer interpret doubt affirm deny believe reject imagine perceive understand concept promise argue declare assert valid justify confirm prove propose evidence utterance logical status ambiguous observe symbolize acknowledge entail summarize premise contradict paradox consistent theory conclude demonstrate discuss define opinion equivalent generalize specify framework abstract concrete unfounded context analyse communicate implicate

The profile results, shown in Figure 2, show that almost 63% of the spontaneously generated items are from the AWL.

In the remainder of this chapter, the foregoing VP analyses of English will be used as a baseline to investigate the question of whether LFP analysis is applicable to a language other than English, and specifically whether a closely related language (French) can be shown to have a zone of lexis resembling the AWL.

WEB VP OUTPUT FOR FILE: *Gedankenexperiment*

K1 Words (1 to 1000):	10	**18.52%**
K2 Words (1001 to 2000):	3	**5.56%**
AWL Words (academic):	34	**62.96%**
Off-List Words:	7	**12.96%**

0-1000 [TTR 10:10] believe declare doubt generalize observe opinion promise propose prove understand
1001-2000 [3:3] argue discuss imagine
AWL [34:34] abstract acknowledge ambiguous analyze communicate concept conclude confirm consistent context contradict define demonstrate deny equivalent evidence framework hypothesize implicate imply infer interpret justify logical perceive reject specify status summarize symbolize theory unfounded utterance valid

OFF LIST [7:7] affirm assert concede concrete entail paradox premise

Figure 2. Web-VP screen output for thought experiment

Note: In this and subsequent screen outputs, K1 and K2 refer to the first and second thousand word lists, respectively, and TTR refers to type-token ratio.

4. Are there AWLs in other languages?

As already mentioned, a study by Hazenberg & Hulstijn (1996) indicated that learners of Dutch would need to know the meanings of 10,000 word families in order to be familiar with 90% of the words in an academic text. However, these researchers did not consider the possibility that Dutch might contain a zone of lexis resembling the English AWL that could foreshorten the learning process. It is not obvious that such a lexical zone would necessarily exist in Dutch. For one thing, the Greco-Latin component of academic discourse is visibly less prominent in Dutch than it is in English. Dutch, like German, has 'traditionally turned to its own resources for enriching the vocabulary' (Stockwell & Minkova 2001: 53). In Dutch we find *natuurkunde* instead of *physics*, *aardrijkskunde* instead of *geography*, *taalkunde* instead of *linguistics,* and so on. Of course, there is nothing to preclude a home-grown AWL in Dutch, German or any other language, which could presumably be located by contrastive corpus analysis.

On the other hand, the mere existence of Greco-Latin items in a language does not necessarily indicate the presence of an AWL. Many of the Greek and Latin AWL items in the English VP output of Figure 2 have cognate counterparts in many other European languages, but these do not necessarily play the same roles, participate in the same genres, or pose the same learning

advantages or difficulties that they do in English. To take a homely example, English speaking children watching a bicycle race in Montreal shout to their heroes 'You can do it!' while their Francophone counterparts shout 'Tu es capable!' The French children are using a word whose equivalent the English children can understand but regard as somewhat formal. In fact, *capable* is an AWL word in English but a high frequency word in French.

The lexical frequency profiles of most languages are not nearly so well known as those of English. One reason is that studies of lexical richness in other languages have often adopted a different methodology, namely type-token analysis, which investigates the amount of lexical repetition in a text rather than the frequency of its lexis with respect to the language at large (e.g., Cossette's 1994 work in French). This methodology has been challenged owing to the way it is influenced by simple text length, which of course is not the case for LFP analysis (see Vermeer, this volume, for an application of LFP analysis to Dutch). Another reason for the limited application of LFP analysis in languages other than English is that the frequency lists that have been developed in these languages have often remained unlemmatized, i.e., do not take the form of word families, so that it is not possible to test their coverage of novel texts (e.g., both Gougenheim, Rivenc, Sauvageot & Michéa's 1967 pre-computational *français fondamental* and Baudot's more recent 1992 computational list). With an unlemmatized list, it is possible for *chat* to be classified as a common word and *chats* as an uncommon word.

In other words, the tools have simply not been available to answer the question: Are there AWLs in other languages? However, in the case of French the situation has recently changed.

5. Recent lexical developments in French

A disparate group of European scholars have recently laid some groundwork for an LFP approach to the description and pedagogy of French. Verlinde & Selva (2001) at Louvain have produced a substantial frequency list of the French language based on a 50-million word collection of recent newspaper texts (*Le Monde* of France and *Le Soir* from Belgium). Glynn Jones (2001) at the Open University in Great Britain has produced a computer program to automatically lemmatize this list, so that parts of it can be run in an LFP-type computer program and their coverage tested with different types of texts.

Goodfellow, Jones & Lamy (2002) have developed a pilot French version of the LFP for pedagogical purposes, and tested it with British students learning French at the Open University. They broke the larger lemmatized frequency list into the familiar 1000, 2000, and (hypothesized) AWL zones, installed these in another web-based version of VP (available at http://iet.open.ac.uk/cgi-bin/vat/vat.html), and used the program to produce lexical profiles of a set of essays produced by learners of French. They looked for correlations between features of these profiles and grades awarded by human raters and found a moderate correlation between proportions of items in the 1001–2000 frequency band and rater scores.

However, the use of these new French lists in a pedagogical analysis may have been somewhat premature, in that no investigation that we know of has examined their reliability across same-genre texts or their differentiation for different-genre texts. Nor was the hypothesized AWL used in this scheme based on analysis of academic texts per se; it was merely the third thousand most common words of general French. No attempt was reported to determine whether these supposed AWL items were indeed more frequent in academic texts than in other genres. The rest of this chapter outlines our attempt to advance this work by micro-testing the reliability, coverage, and genre characteristics of these potentially useful French lists. But before that a certain amount of preliminary work had to be done on the lists themselves.

The three French frequency lists used in the Goodfellow *et al.* (2002) experiment were generously shared with us by the researchers. We then incorporated these lists into a French web-based version of VP, to be known as *Vocabprofil*, which like its English counterpart allows full inspection of classifications into frequency ranges (see the screen output in Figure 2). A large number and variety of texts were run through the program and output profiles were inspected in detail for anomalies. The output checking was performed by French native-speaker research assistants, but also by general users of the public website who emailed their queries. These users reported both inconsistencies in lemmatization (e.g., *participer* was listed as a first 1000 item while *participant* was listed as second) and simple misclassifications (several French speakers found it odd to see *calme* listed as an off-list or uncommon word). Items flagged as potential misclassifications were checked against a second French frequency list (Baudot 1992), and about 200 items were either reclassified or added to the three lists over a six-month testing period. The size of the lists remained almost identical throughout this process with the number of additions (e.g., *calme*) roughly equaling the number of deletions through

reclassification (e.g., *participant* removed as a second thousand entry and reclassified under *participer* in the first 1000).

Our decisions about which words to count as members of the same family were taken in the spirit of the work done in English by Bauer & Nation (1993), where frequency, regularity and transparency of inflection and affixation were the main criteria for family inclusion (e.g., *participant* is an obvious relation of *participer* for anyone likely to be reading a text including these items). It should be noted, however, that the algorithmic lemmatization procedure that had previously been applied by Jones (2001) to these French lists was intended to be exhaustive, such that some low frequency affixes were attached to high frequency base words; hence, some forms that learners might not easily recognize were categorized as frequent. (How readily would a beginning learner of French recognize *échappassiez* as a form of *échapper*?) The word families that emerge from this lemmatization procedure are truly enormous, particularly in the case of verbs with their many suffixes. These French lists, if ever intended for learners, would eventually need to be reworked along the lines of Bauer & Nation's (1993) procedure. An interim solution to their pedagogicalization is suggested in the conclusion.

Once satisfied that the English and French lists were more or less comparable, as signaled by a cessation of anomaly reports from users and research assistants, we shifted our focus from the lists themselves to the comparison of the lexical frequency profiles of English and French that the lists made possible.

6. Preliminary investigation of French profiles

A large bilingual and multi-generic corpus analysis such as that provided by Coxhead (2000) for English will eventually be required to complete the investigation we are beginning. In this preliminary investigation, we are merely applying our new word lists to French texts in order to test the feasibility of LFP analysis and get a sense of its pedagogical potential. Our methodology is to collect a small bilingual corpus of medium-sized original and translated texts of distinct genres, run these piecemeal through both *Vocabprofile* and *Vocabprofil*, and compare the results. Through this we hope to answer specific questions about the coverage of the lists, and the reliability and genre specificity of the profiles they provide, always with a view to answering the larger question of whether there is anything resembling an AWL in French. We begin with the issue of reliable coverage.

Question 1: Do the new French lists produce reliable coverage profiles within a text genre?

To answer this question, we had our graduate students select and download 100 typical online news texts of between 500 and 1000 words on political topics from different parts of the Francophone world, and submit these one by one to VP analysis using the Internet version of the program running the new frequency lists described above. Proper nouns were not eliminated or otherwise treated, on the assumption that most political news stories would carry similar proportions of these. By analyzing the texts separately, rather than as a single unit, we were able to calculate a measure of variability for frequency categories across the collection. Profiles across these texts proved remarkably consistent, displaying very low degrees of within-level variance. The results of eight of these analyses can be seen in Table 4. The main point of interest is that the standard deviations across the 1000, 2000, and AWL zones are small, even smaller than the figures for English in Tables 2 and 3 above. All adjacent pairs of texts were subject to *chi*-square tests of comparison; none of the profile differences proved to be statistically significant.

Table 4. Consistent profiles within news texts (coverage percentages)

Paper	Topic	1000	2000	AWL	1K + 2K + AWL	Off-list
Le Devoir	CBC coupures	79.65	9.29	1.33	90.27	9.73
La Presse	CBC coupures	82.78	7.50	0.96	91.23	8.77
Le Devoir	Abandon scolaire	78.00	10.85	2.17	91.01	8.99
La Presse	Abandon scolaire	77.15	9.18	2.34	88.67	11.33
Le Devoir	Bush & Irak	75.10	8.54	1.88	85.52	14.48
La Presse	Bush & Irak	77.68	7.40	3.20	88.29	11.71
Le Monde	Bush & Irak	75.99	8.59	1.98	86.56	13.44
Figaro	Bush & Irak	74.75	7.43	1.82	83.99	16.01
M		77.64	8.60	1.96		11.81
SD		2.63	1.19	0.67		2.65

With the new lists apparently able to produce reliable profiles, at least for this type of text, we next turn to the more interesting question of their coverage. Standard deviations will continue to be provided in the analyses to follow, as a continuing reliability check for other types of texts.

Question 2: Do the French lists provide similar text coverage to their English counterparts?

To answer this question, we used French translations of the 18 *Reader's Digest* texts already seen for English above (translated by bilingual research assistants in Montreal[2]). Figure 3 provides a sample translation of one of the texts. The question of interest was whether the French lists as developed to date would produce consistent coverage figures across texts for each of three frequency zones (1000, 2000, AWL), as indicated by small deviations from the mean coverage figures for the 18 texts.

The translated texts were fed through *Vocabprofil* piecemeal. The resulting profile means and standard deviations are shown in Table 5a, and comparison figures for the English profiles of the same texts just below in Table 5b.

The first thing to note in these results is that once again, the experimental French lists provide consistent amounts of coverage across texts on disparate

Traditional methods of teaching are no longer enough in this technological world. Currently there are more than 100,000 computers in schoolrooms in the United States. Students, mediocre and bright alike, from the first grade through high school, not only are not intimidated by computers, but have become enthusiastic users.	Les méthodes traditionnelles d'enseignement ne suffisent plus dans ce monde de technologie. Présentement, on compte plus de 100 000 ordinateurs dans les salles de classe aux États-Unis. Les étudiants, moyens et brillants pareils, de la première année jusqu'à la fin du secondaire, sont non seulement peu intimidés par les ordinateurs mais sont même devenus des utilisateurs enthousiastes.
Children are very good at using computers in their school curriculum. A music student can program musical notes so that the computer will play Beethoven or the Beatles. In a biology class, the computer can produce a picture of the complex actions of the body's organs, thus enabling today's students to understand human biology more deeply. A nuclear reactor is no longer a puzzle to students who can see its workings in minute detail on a computer. In Wisconsin, the Chippewa Indians are studying their ancient and almost forgotten language with the aid of a computer.	Les enfants sont très doués pour ce qui est d'utiliser les ordinateurs dans leur curriculum scolaire. Un étudiant en musique peut programmer des notes pour que l'ordinateur joue Beethoven ou les Beatles. Dans un cours de biologie, l'ordinateur peut produire une image du fonctionnement complexe des organes du corps, permettant ainsi à l'étudiant de comprendre plus en profondeur les principes de la biologie humaine. Un réacteur nucléaire n'a plus de mystères pour les étudiants qui peuvent observer son fonctionnement en détails sur l'ordinateur. Dans le Wisconsin, les indiens Chippewa s'en servent pour étudier leur langue, ancienne et presque oubliée.
The simplest computers aid the handicapped, who learn more rapidly from the computer than from humans. Once a source of irritation, practice and exercises on the computer are now helping children to learn because the machine responds to correct answers with praise and to incorrect answers with sad faces and even an occasional tear. - 198 words	Les ordinateurs les plus simples aident les personnes handicapées qui apprennent plus rapidement d'une machine que d'une autre personne. Jadis une source d'irritation, la pratique et les exercices sur l'ordinateur aident maintenant les enfants à apprendre parce que la machine complimente les bonnes réponses et présente des visages tristes et même parfois une larme pour les mauvaises réponses. - 214 words

Figure 3. Sample translation: Learning with computers/Apprentissage par ordinateur

Table 5a. French VP profiles for popular expository texts (coverage percentages)

TEXT	1000	2000	1000 + 2000	AWL	1K + 2K + AWL	Off-list
Audubon	71.51	8.72	80.23	1.16	81.39	18.60
Computers	78.85	10.13	88.98	3.52	92.50	7.49
Drugs	74.24	6.11	80.35	2.18	82.53	17.47
Earthquakes	73.33	10.26	83.59	4.62	88.21	11.79
Dieting	83.21	6.87	90.08	3.82	93.90	6.11
Gas	74.07	14.07	88.14	5.93	94.07	5.93
Olympics	75.29	6.32	81.61	2.30	83.91	16.09
Origins of Life	70.80	8.41	79.21	3.98	83.19	16.81
Plague	76.29	5.15	81.44	4.12	85.56	14.43
Salt	76.95	5.47	82.42	2.73	85.15	14.85
Stage fright	76.08	7.84	83.92	2.35	86.27	13.73
Teenagers	81.14	8.00	89.14	2.86	92.00	8.00
Tennis	75.95	9.28	85.23	1.69	86.92	13.08
Toledo	71.78	7.92	79.70	2.97	82.67	17.33
Vitamins	70.94	8.37	79.31	3.94	83.25	16.75
Volcanoes	78.74	5.91	84.65	3.54	88.19	11.81
Warm-Up	79.02	8.93	87.95	4.46	92.41	7.59
M	75.78	8.10	83.88	3.30	87.18	12.82
SD	3.61	2.18	3.78	1.20	4.32	4.32

Table 5b. Comparison of French and English profiles (coverage percentages).

	1000		2000		1000 + 2000		AWL		1K + 2K + AWL	
	M	SD	M	SD	M	SD	M	SD	M	SD
English	74.95	5.73	6.32	2.62	81.26	4.47	5.56	1.99	86.83	4.56
French	75.78	3.61	8.10	2.18	83.88	3.78	3.30	1.20	87.18	4.32

topics. Across the 17 texts, the first 1000 most frequent French words account for about three quarters of the lexis of each text, while the second thousand and AWL add approximately 8% and 3%, respectively. Standard deviations from these means are small and consistently lower than their English counterparts, suggesting even greater reliability. The second thing to note is that the overall coverage provided by the English and French lists appears to be similar across the two languages, within a percentage point (86.83% for English and 87.18% for French) for the three lists combined. However, there is a hint of a difference in the components that are providing the coverage in the two languages. The French second thousand list appears to account for more items than its English counterpart, the English AWL for more items than its French counterpart.

This apparent asymmetry will be further explored with texts of a more academic character below.

Question 3: Is there an AWL in French?

To launch an investigation of this question, we first located a set of nine translated medical texts of equivalent size (about 1000 words each) on the website of the *Canadian Medical Association Journal* (CMAJ) and submitted them to piecemeal bilingual VP analysis. Some of these had been translated from English to French, and others from French to English. The results are shown in Table 6a and summarized in Table 6b. It should be noted that medical texts are typically dense in domain-specific terms, which may explain why the off-list component here is high (about 20% of items are not accounted for by the three lists, as opposed to the usual 10 or so). Still, some interesting points of comparison emerge in the more frequent zones (see Table 6b). The first thing to note in the results is that the experimental French AWL does not pull any more weight in these medical texts (only 3.57%) than it did in the

Table 6a. Profiles of medical texts, translated but in no consistent direction (coverage percentages).

Text	English			French		
	1000	2000	AWL	1000	2000	AWL
Dental 1	59.84	7.10	9.56	66.39	11.55	5.57
Dental 2	68.43	9.85	8.37	72.46	7.66	3.79
Breast self exam	65.67	8.50	12.44	73.31	9.37	3.91
Preventive care	63.13	8.94	12.85	72.41	11.35	3.30
Mammography	61.90	9.52	15.00	72.70	10.43	3.13
Lymphedema	57.14	8.12	12.04	69.40	8.62	3.02
Biopsy	54.50	7.63	12.26	67.13	11.95	2.99
Child abuse	60.03	14.46	15.31	76.19	11.19	3.21
Hyperhomocystein-emia	60.84	8.16	10.72	71.04	8.27	3.24
M	61.28	9.14	12.06	71.23	10.04	3.57
SD	4.20	2.17	2.28	3.11	1.60	0.81

Table 6b. Means comparisons for Table 6a (coverage percentages).

	1000		2000		1K + 2K		AWL		1K + 2K + AWL	
	M	SD	M	SD	M	SD	M	SD	M	SD
English	61.28	4.20	9.14	2.17	70.42	5.07	12.06	2.28	82.48	5.54
French	71.23	3.11	10.04	1.60	81.27	3.17	3.57	0.81	84.84	3.02

Readers' Digest texts (3.30%). The second thing is that again the French first and second thousand lists are doing more work (covering 81.27% of running items) than the corresponding English lists (covering 70.42%). It appears that in English, the AWL is needed to achieve the same coverage provided by just the first and second thousand French lists. Our comparison so far suggests that in terms of coverage power, the French first and second thousand lists are stronger than their English counterparts, and that the (hypothesized) French AWL is weak.

We next sought a confirmation of these interesting differences between the French and English lists by working with a different set of texts, this time from a domain with fewer off-list items than the medical texts. We found a set of ten 2000-word translated speeches from the European Parliament which had two characteristics that made them a good test of the French lists. First, they were originally written in French (for the most part) unlike the translations and mixtures of translated and original writing used earlier. Secondly, the fact that the English versions were unusually AWL-rich meant that there was ample scope for any French AWL to emerge. These French and English texts were downloaded from the Internet and run separately through their respective VPs. Summary results are shown in Table 7.

The main thing to note about this second comparison is that, again, English needs its AWL to approach the 90% coverage mark (1000 + 2000 + AWL = 89.43%), while French can get there using its high frequency words alone (1000 + 2000 = 88.32%).

It is tempting to conclude that the French second thousand list is doing some of the work that the AWL does in English. Such a conclusion is consistent with the Goodfellow *et al.* (2002) finding mentioned above that a higher proportion of second thousand-level items tends to predict a higher score on a composition (a role normally played by the AWL in English). This brings us to the central question of our investigation.

Table 7. Mean profiles for 10 EU speeches, translated mainly French to English (coverage percentages).

	1000		2000		1K + 2K		AWL		1K + 2K + AWL	
	M	SD	M	SD	M	SD	M	SD	M	SD
English	79.21	2.17	3.42	0.56	82.63	2.67	6.81	1.74	89.43	1.02
French	78.36	0.92	10.30	1.32	88.32	0.52	2.02	0.59	90.34	0.34

Question 4: Is there room for an AWL in French?

The purpose of identifying the English AWL was to offer university-bound learners of English a shortcut to achieving 90% known word coverage of texts, a target that can otherwise be achieved only when 6000 or more word families have been acquired through naturalistic exposure (see Table 1). From the preliminary data presented here, it appears that learners of French can reach the same level of coverage of an academic text (or any other kind of text) when they have acquired knowledge of the meanings of just the 2000 most common French word families. If we can assume that the remaining 10% of items (or more in some domains such as medicine) comprise proper nouns, domain specific items, and a few other low frequency words in roughly the same proportions as English, then the best guess at this point would have to be that there is no room left in French for an AWL.

The lack of an AWL does not imply that French has fewer resources for academic discourse; rather, it appears to conduct this discourse without the need of a distinct body of words different from those used in everyday life. After all, it is English that is typologically special in having two distinct strands interwoven in its lexicon, the Anglo-Saxon and the Greco-Latin, from whence the phenomenon of the AWL. But in English, or in any language, common words can be used in more or less common ways. Both *To be or not to be* and *Je pense donc je suis* convey complex ideas using common vocabulary, and it is perfectly conceivable that for English this is an option while for French it is the norm. A further test of this idea is presented below.

Question 5: Are the academic words of English the common words of French?

Like all of our questions, this one can receive a definite answer only through large corpus analysis, but in the meantime the methodology of the *Gedanken-experiment* developed in our introduction may provide a hint. In this experiment we determined that the lexis of argument and discourse in English comprised mainly AWL items (*hypothesize, imply, infer,* etc.). If French has an AWL somewhere beyond its first 2000 words, then presumably the French versions of these words would be in it. To investigate the frequency status of these words in French, we translated all plausible French cognates of the 54 original words and checked them with native speakers. The list eventually submitted to analysis included only obvious cognates (e.g., *context* and *contexte*); unclear cases were eliminated (e.g., *summarize*), leaving the following list of 42 words:

impliquer hypothèse inférer interprétation douter affirmer nier rejeter imaginer percevoir concept promesse déclarer assertion valide justifier confirmer prouver proposer évidence logique statu ambigu observer symbolise contradiction paradoxe consistent théorie conclusion démonstration discussion définir opinion équivalent généraliser spécifier abstrait concrète contexte analyser communication

This word-set was run through *Vocabprofil*, and the output is shown in Figure 4.

In English, 63% of these words are AWL items; in French almost 56% (39.53 + 16.28) are first 2000 items. A large role for the 2000 list emerged again when all 570 headwords of the English AWL were passed through Altavista's *Babel Fish* translation routine (http://babelfish.altavista.com/) and the French translation equivalents were submitted to *Vocabprofil* analysis. About 58% of these translations proved to be on the 2000 most frequent list, with 26.69% on the first 1000 list and 30.90% on the second. Both the *Gedankenexperiment* and the translation exercise point to the same conclusion: The French second thousand list seems again to be a main repository of academic words (or, as becomes increasingly apparent, of words that are commonly used but can also render academic service). Of course, it is not the only repository of such words; a further 46% of the AWL equivalents and a similar proportion of the *Gedankenexperiment* items are from less frequent zones (e.g., *abstrait, ambigu, concrète; assertion, concept, douter)* but these are unlikely to constitute a coherent lexical zone comparable in size and function to the AWL, given that these words must share a 10% space

WEB VP OUTPUT FOR FILE: *Gedankenexperiment* 2

Mots K1 (1 à 1000):	7	**16.28%**
Mots K2 (1001 à 2000):	17	**39.53%**
Mots K3 (2001 à 3000):	11	**25.58%**
Mots Off-List:	8	**18.60%**

0-1000 [7 types 7 tokens]: communication discussion déclarer imaginer opinion proposer

1001-2000 [17 types 17 tokens]: affirmer analyser conclusion confirmer consistent contexte définir hypothèse impliquer interprétation justifier logique observer percevoir prouver rejeter évidence

2001-3000 [11 types 11 tokens]: abstrait ambigu concrète contradiction démonstration généraliser nier paradoxe symbolise théorie valide

Off list [8 types 8 tokens]: assertion concept douter inférer promesse spécifier statu équivalent

Figure 4. French Cognates Gedankenexperiment profile

with proper nouns and domain-specific items. (Note that in the screen picture, the name K3, for third thousand, is now used to designate what the originators of these French lists hypothesized might be an AWL.)

Having answered the main questions to the extent allowed by the scale of our present investigation, we return to one trailing matter — the question of profiles and genres.

Question 6: Do French genres have distinct LFPs?

The evidence needed to answer this question has already appeared in the answers to questions above. With the exception of medical texts, the French profiles are not distinct across genres (according to *chi*-square comparisons). For example, first 1000 mean coverage percentages are very close across text types, with newspaper texts at 77.64%, *Readers' Digest* texts at 75.78%, and EU speeches at 78.36% (Tables 4, 5a, 7, respectively). By contrast, English counterpart profiles are quite distinct, largely owing to the different role of the AWL in the different genres.

7. Conclusion

First, it has been gratifying to work with the French lexical resources recently made available by Selva, Verlinde, Goodfellow and their colleagues. These resources have long been needed in the pedagogical study of French, and we hope that we have made a contribution. The *Vocabprofil* website is receiving a lot of visitors so perhaps we have at least publicized the LFP approach to the analysis of French.

On the general question of the comparative lexical distributions of English and French, the evidence we have gathered suggests that these distributions may be somewhat different. They are similar in that both English and French (and probably any language) use their most frequent 1000 or 2000 words quite heavily, but different in that French seems to use its frequent words even more heavily than English does. On the specific question of whether there is room for an AWL in French, the provisional answer appears to be that there is not. In almost all of the cases we examined, it appears that the goal of achieving 90% text coverage can be met by mastering the common vocabulary of French

Vocabulary has traditionally not been considered the most important thing to emphasize in the teaching of French, and this is an intuition that appears to have some basis. This is not to say that learning the academic

vocabulary of French is easy, because the challenge of learning academic uses of common words is probably just as great as learning new academic words. But it is to say that the acquisition process is able to proceed on a naturalistic basis for learners of French as it is not for learners of English.

Any naturalistic learning can nevertheless be made more efficient, possibly by using some of the resources and technologies we have discussed. For example, teachers could use *Vocabprofil* to tailor reading materials to their learners' level of lexical development. Independent learners might wish to peruse the lists themselves and check for any gaps in their knowledge. This idea has proved to be popular with English-learning users of the *Lexical Tutor* website, who can click on 1000, 2000, or AWL words to hear them pronounced and see them contextualized in concordances and dictionaries. The French lists, however,

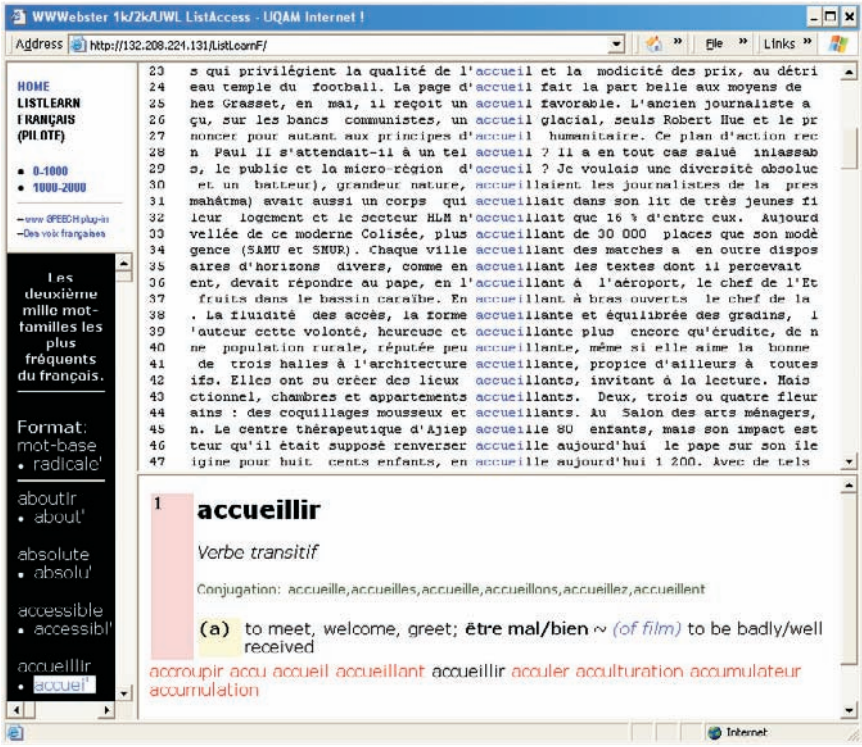

Figure 5. Pedagogical version of French list.

Note: The word lists are hyper-linked to a 1 million word corpus of *Le Monde* newspaper texts, provided by Thierry Selva in Louvain (upper right frame); French-English Dictionary was developed by Neil Coffey in Oxford (lower right frame).

with full lemmatization are simply too large for pedagogical presentation (the 2000 list is dozens of pages) let alone for Internet delivery. A solution to this problem was to de-lemmatize the verbs and replace them with the infinitive plus a homemade radical that would generate all the contextualized forms of the word in a French corpus. Figure 5 illustrates the case of the verb *accueillir* and the radical *accuei'* used to search for the corpus for forms beginning with this string.

Finally, on a more general note, it is interesting to contemplate the possibility that each language may have its own lexical shape, each entailing different acquisition strategies for learners and different teaching strategies (not to mention different tutorial computer programs). From the best evidence we have right now, it seems that reading academic texts in Dutch, English, and French may require widely different amounts of lexical knowledge and different kinds of lexical skills. There can be no doubt that this idea is interesting enough to warrant further research.

Notes

1. This and many other factual claims made in this chapter can be tested on the first author's *Lexical Tutor* website (Cobb, online).

2. Translations were generously provided by Norman Segalowitz and his research team in the Psychology Department of Concordia University in Montreal.

References

Babel Fish Translation [Online: available at http://babelfish.altavista.com/].

Baker, C. L. and McCarthy, J. (eds). 1981. *The Logical Problem of Language Acquisition.* Cambridge MA: MIT Press.

Baudot, J. 1992. *Fréquence d'utilisation des mots en français écrit contemporain.* Montréal: Les Presses de l'Université de Montréal.

Bauer, L. and Nation, P. 1993. "Word families". *International Journal of Lexicography* 6: 253–279.

Cobb, T. The compleat lexical tutor for data-driven language learning on the web. [Online: available at http://www.lextutor.ca].

Corson, D. 1985. *The Lexical Bar.* Oxford: Pergamon Press.

Corson, D. 1997. "The learning and use of academic English words". *Language Learning* 47: 671–718.

Cossette, A. 1994. *La richesse lexicale et sa mesure.* Paris: H. Champion.

Coxhead, A. 2000. "A new academic word list". *TESOL Quarterly* 34: 213–238.

Coxhead, A. The academic word list [Online: available at http://www.vuw.ac.nz/lals/div1/awl/].

Francis, W. N. and Kucera, H. 1979. *A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers.* Department of Linguistics, Brown University.

Goodfellow, R., Jones, G. and Lamy, M.-N. 2002. "Assessing learners' writing using Lexical Frequency Profile". *ReCALL* 14: 129–142.

Gold, E. M. 1967. "Language identification in the limit". *Information & Control* 10: 447–474.

Gougenheim, G., Rivenc, P., Sauvageot, A. and Michéa, R. 1967. *L'élaboration du français fondamental (Ier degré).* Paris: Didier.

Hazenberg, S. and Hulstijn, J. 1996. "Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation". *Applied Linguistics* 17: 145–163.

Jones, G. 2001. "Compiling French word frequency list for the VAT: A feasibility study". [On-line] Working paper. [Online : available at http://www.er.uqam.ca/nobel/r21270/cgi-bin/F_webfreqs/glynn_jones.html].

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

Nagy, W. E. and Herman, P. A. 1987. "Depth and breadth of vocabulary knowledge: Implications for acquisition and instruction". In *The nature of vocabulary acquisition*, M. G. McKeown & M. E. Curtis (eds), 19–35. Hillsdale, NJ: Erlbaum.

Nation, I. S. P. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Olsen, D. 1992. The World on Paper: *The Conceptual and Cognitive Implications of Writing and Reading.* Cambridge: Cambridge University Press.

Pinker, S. 1995. *The Language Instinct.* New York: Harper-Collins.

Stockwell, R. and Minkova, D. 2001. *English Words: History and Structure.* New York: Cambridge University Press.

Sutarsyah, C., Nation, P. and Kennedy, G. 1994. "How useful is EAP vocabulary for ESP? A corpus based study". *RELC Journal* 25: 34–50.

Verlinde, S. and Selva , T. 2001. "Corpus-based vs. intuition-based lexicography: Defining a word list for a French learners' dictionary". In *Proceedings of the Corpus Linguistics 2001 conference*, Lancaster University, 594–598. [Online: available at http://www.kuleuven.ac.be/ilt/grelep/publicat/verlinde.pdf].

Xue, G. and Nation, I. S. P. 1984. "A university word list". *Language Learning & Communication* 3: 215–229.