Bilingualism: Language and Cognition 11 (1), 2008, 133–146 © 2008 Cambridge University Press doi:10.1017/S1366728907003264

Words that second language learners are likely to hear, read, and use*

DOUGLAS J. DAVIDSON
F. C. Donders Centre for Cognitive Neuroimaging
Max Planck Institute for Psycholinguistics
PETER INDEFREY
F. C. Donders Centre for Cognitive Neuroimaging
Max Planck Institute for Psycholinguistics
MARIANNE GULLBERG
Max Planck Institute for Psycholinguistics

In the present study, we explore whether multiple data sources may be more effective than single sources at predicting the words that language learners are likely to know. Second language researchers have hypothesized that there is a relationship between word frequency and the likelihood that words will be encountered or used by second language learners, but it is not yet clear how this relationship should be effectively measured. An analysis of word frequency measures showed that spoken language frequency alone may predict the occurrence of words in learner textbooks, but that multiple corpora as well as textbook status can improve predictions of learner usage.

Research on vocabulary knowledge in second language (L2) learners has for the most part concentrated on developing measures to assess the proficiency or vocabulary size of individual learners. This study addresses a different but related question: how does an experimenter or researcher choose words that learners of an L2 are likely to know during acquisition? Experimenters may need to select words likely to be known by learners to test non-lexical aspects of L2 knowledge. Similarly, when constructing materials for students, instructors must in some cases choose materials so that learners are likely to already know the words in a lesson, so that other, non-lexical aspects of the lesson are conveyed effectively. For child language learners there are vocabulary tests such as the Peabody Vocabulary Test (Dunn and Dunn, 1997) available with estimates regarding which words are known at a particular age. No such tests exist for second language learners. Conversely, researchers will sometimes require words that learners are UNLIKELY to know, if a study is designed to examine how learners process or learn unfamiliar words. On what basis should these choices be made?

A straightforward answer seems to be that words should be selected on the basis of their lexical frequency. Word frequency is an important factor in second language acquisition (SLA), as learners tend to recall (Laufer, Elder, Hill and Congdon, 2004) and recognize (Brown, 1993) words that are more common in a given language better than less frequent words. Based on this relationship between lexical statistics and vocabulary knowledge, it is generally accepted that knowing less frequent words is indicative of having a larger vocabulary or, as Meara (2005, p. 32) puts it, "[the] use of frequency bands to characterize vocabulary is a fairly standard practice in L2 vocabulary studies". However, the available word frequency data are based on samples from different sources such as newspapers, books and conversation, and this leads to the problem of whether estimates from all sources are equally appropriate for the population of second language learners. Samples of language use from which a set of words is selected may be biased in the sense that some of the properties of the sample words may not represent accurately the properties of the entire language from which it is sampled while other properties may in fact be represented accurately. In newspaper text, for example, words related to certain topics, such as politics, are typically overrepresented. When constructing materials for experiments or teaching purposes, the bias of the corpora should not enter as a bias in the materials.

More importantly, the language to which learners are exposed and therefore the vocabulary they acquire, is itself biased. Even an ideal sample that represents the entire language correctly would not necessarily yield a frequency estimate that optimally predicts the vocabulary of learners. Most clearly this is the case for classroom learning where the frequency of words that learners

* Arna van Doorn assembled the vocabulary lists from the three Dutch textbooks. The Max Planck Institute for Psycholinguistics provided access to the CELEX, the CGN, and the ESF corpora. The analysis was conducted using *R* (R Development Team, 2005), and the *stats* (R Development Team, 2005) and *MASS* (Venables & Ripley, 2002) libraries. This research was supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). We would also like to thank two anonymous reviewers for useful suggestions, and Jan Hulstijn for providing helpful comments and references for textbook vocabulary selection including, in addition to those cited in the text, Hazenberg (1994) and Sciarone (1979).

Address for correspondence:
Doug Davidson, F. C. Donders Centre for Cognitive Neuroimaging, P.O. Box 9101, 6500 HB Nijmegen, The Netherlands
doug.davidson@fcdonders.ru.nl

encounter does not reflect the entire language, but rather the teaching materials. It is known that words that are more commonly encountered in school are more likely to be known on vocabulary tests. This has been demonstrated in experimental studies of vocabulary acquisition in reading (Pitts, White and Krashen, 1989; Day, Omura and Hiramatsu, 1991; Dupuy and Krashen, 1993; Hulstijn, Hollander and Greidanus, 1996; Rott, 1999), as well as through an empirical demonstration that the frequency of words presented in classroom materials predicts the likelihood of a correct answer on tests of vocabulary (Vermeer, 2001), and demonstrations that textbook appearance can be used as a predictor for word recognition ability (Van Gelderen et al., 2004; Fukkink, Hulstijn and Simis, 2005).

The problem of the classroom bias suggests the possibility that the occurrence of words in teaching materials such as textbooks may in fact be more informative with respect to early vocabulary than corpus frequencies, at least for classroom learners. Biases of the language that untutored learners are exposed to are obviously much more variable depending on the context of language input. Assuming, however, that textbook authors have a good intuition about the kind of vocabulary all language learners, irrespective of their learning circumstances, need, it is possible that textbook occurrence may be a good predictor for the early vocabulary of untutored early learners as well.

The goal of the present paper is to assess whether multiple data sources may be more effective than single data sources to predict the words that language learners are likely to know. We will first address the question of how word frequencies estimated from corpora are related to textbook vocabulary. Do frequency estimates from text corpora predict textbook occurrence with equal effectiveness as those from spoken corpora? In a second step, we will describe an estimation procedure using several data sources at once. We will assess which combination of word frequency data from speech and text corpora and occurrence in learner textbooks best predicts the occurrence of nouns in a corpus of untutored L2 learners' speech. The goal is to find nouns that learners are likely to know across a broad range of settings both inside and outside the classroom.

Method

In the analyses reported below, we investigate the relationship between three types of information: (1) Estimates of noun form frequency taken from spoken and written corpora representing a wide variety of word usage, (2) vocabulary lists taken from adult foreign language learning textbooks, and (3) estimates of word usage taken from a small sample of untutored language learner speech based on a film description task. The goal for the analysis is to characterize the relationship between different corpus frequencies and to evaluate how well different corpus frequencies, learner textbook occurrence, and different combinations of these data sources predict actual learner production. Based on the best-predicting combination, the analysis should provide a list of words that second language learners are likely to know and use.

Word frequency lists typically contain a list of word forms paired with the number of instances of the word form in a corpus sample. Word frequency distributions are heavily skewed with the majority of word types having very low frequencies. This skew is a problem for researchers who wish to select common words because the numerical value of any frequency estimate will depend heavily on the size of the sample of speech or text. This makes it difficult to choose a fixed frequency count as a threshold for common words. In the example lists described below, we selected from word form frequency lists words with frequencies above the log average frequency based on words with frequencies > 1 in order to reduce the influence of very low frequency words. Rank information could be used in place of frequency counts, but frequencies potentially contain information that is lost in ranks. There may be a large difference between two rank-adjacent words in frequency, yet another pair of rank-adjacent words may have a small difference in frequency.

Two corpora were used for this study, the CGN (Corpus Gesproken Nederlands or the Spoken Dutch Corpus) and CELEX. The CGN word frequencies (Oostdijk, 2000; based on version 1.0 of the Spoken Dutch Corpus) were based on nouns taken from the CGN Lexicon portion of the corpus (Dutch and Flemish, comprising 181,579 word forms). Frequencies were scaled to frequencies per million. Only the single word-form frequencies were included in the analysis presented (the corpus includes information for multiple-word forms (phrases) in a separate lexicon). The corpus consists of samples of conversational and non-conversational speech including telephone conversations, debates, broadcast news reports and commentary, speeches, and lectures (a detailed description of the components is available in the documentation of the distributed corpus).

The CELEX word frequencies (Baayen, Piepenbrock and Gulikers, 1995) were based on the word form and word lemma frequencies included in the CELEX database. CELEX frequencies for Dutch are based on a sample of fiction (30%) and non-fiction (70%) books published between 1970 and 1988. The word frequency estimates do not include material from newspapers, magazines, children's books, or textbooks.

The words from the textbooks were taken from three modern introductory language textbooks for Dutch (selected because they were conveniently available): *Taal vitaal: Nederlands voor beginners* (Schneider-Broekmans, 2000), *Dutch for self-study* (van Kampen

and Stumpel, 2002), and *Colloquial Dutch: The complete course for beginners* (Donaldson, 1996). All three textbooks are aimed at beginning learners of Dutch and include word lists. For the word form-based analyses described below we expanded the set of words from the textbook word lists (which only provide citation forms) by adding to it all noun forms of the lemma. If the word *kat* "cat" appeared in the textbook list, for example, then *katten* "cats" and *katje* "cat, diminutive", and *katjes* "cats, diminutive" were added to the expanded textbook list.

Textbook wordlists, including the textbooks chosen in the present study, are likely to be selected based on word frequency (Bossers, 1996). Beginner textbooks concerning Dutch commonly use a wordlist compiled in De Kleijn and Nieuwborg (1983), based on frequency estimates produced in Uit den Bogaart (1975). While it would therefore be expected that word frequency would predict whether a word will be found on a textbook wordlist, earlier word frequency estimates have largely been based on text, and not speech, as large speech corpora have only become available recently. It is therefore not clear whether both types of frequency estimates would contribute to the prediction. In the present paper, the relative predictive power of text versus speech word frequency estimates is compared.

The learner production data consisted of words taken from a film-re-telling task collected as part of the European Science Foundation (ESF) Second Language Learner corpus (see Perdue, 1984; 1993 for details). The film re-tellings covered restricted portions of Chaplin's film "Modern Times". The corpus consisted of the film descriptions of four Moroccan and four Turkish immigrant learners of Dutch on two measurement occasions (two re-tellings of the same film) within the first year of learning Dutch. As such, the film re-tellings represent a sample of L2 production in which the same participants produced descriptions of a singe film, in many ways like an experimental setup in which different participants produce sentences to an experimental stimulus set. The learners in this sample were untutored, that is, learners without exposure to classroom teaching and textbook materials.

For the nouns that were found in the Dutch learner textbooks, we estimated how well the spoken and written Dutch corpora (CGN and CELEX) frequencies predicted the status of a word as a textbook vocabulary entry (yes or no) using logistic regression. We use logistic regression to determine how strongly spoken and textual frequencies are related to the propensity to be found in an introductory Dutch textbook wordlist. For the analyses presented below, we compared model fits using deviance (p-values are reported for 1-df Chi-square comparisons), as well as the Akaike Information Criterion (AIC; Akaike, 1974) using stepwise comparison (both forward and backward) of model terms. AIC is a criterion for selecting between different statistical models (such as regression models), and is used widely in regression analysis. It is defined as –2L –2m, where L is the log likelihood of the model under consideration, and m is the number of parameters in that model. The general benefit of applying AIC is that it takes into account both the goodness of fit (the likelihood of the model) as well as the number of parameters in the fitted model (the model complexity); smaller AIC values indicate a better-fitting model. Normalized differences in AIC values between competing fits can be interpreted as proportional likelihoods. To ensure that the results of our analyses did not depend on assumptions about the exact nature of the entries in the learners' mental lexicon (e.g. whether or not unanalyzed inflected forms are listed), we estimated the fit based on word forms as well as on lemmas (summed frequencies of all forms of a word).

Results

Corpus predictions of textbook vocabulary status

Figure 1 shows the empirical proportion of word forms present in the textbooks and the predicted proportions for the models including CGN or CELEX for the range of scaled log CGN and CELEX form frequencies. This plot shows how the more common words in either of the corpora are more likely to be found in the textbooks, especially for the range of standardized frequency from 1.0 to 3.0. A model that included a positive term for CGN log frequency was the best-fit model (AIC = 4541), compared to CELEX alone (AIC = 5039). The model that included both CELEX and CGN (AIC = 4540) did not significantly improve the fit over the CGN-alone model ($p = 0.1$). The parameters for the best-fitting model were Proportion_Textbook = –0.7039 + 1.1959 * Frequency_CGN.

An additional analysis was conducted by using CELEX and CGN lemma frequencies rather than form frequencies. Figure 2 shows the empirical proportions of lemmas present in the textbooks and the predicted proportions based on the models including CGN or CELEX for the range of scaled log CGN and CELEX lemma frequencies. Note that given that there are more low frequency than high frequency words, the parameters of the best fitting models will be such that the predictions for the low frequency ranges are better. It then appears that predictions based on CELEX underestimate textbook appearance for higher frequency items. Results of this analysis were similar to the previous analysis. The best fitting model included only CGN log lemma frequency (AIC = 3057), compared to CELEX alone (AIC = 3390; $p < 0.001$), and the conjunction of CGN and CELEX frequencies (AIC = 3062) did not improve the fit ($p = 0.48$) over CGN alone. The parameters for the best-fitting

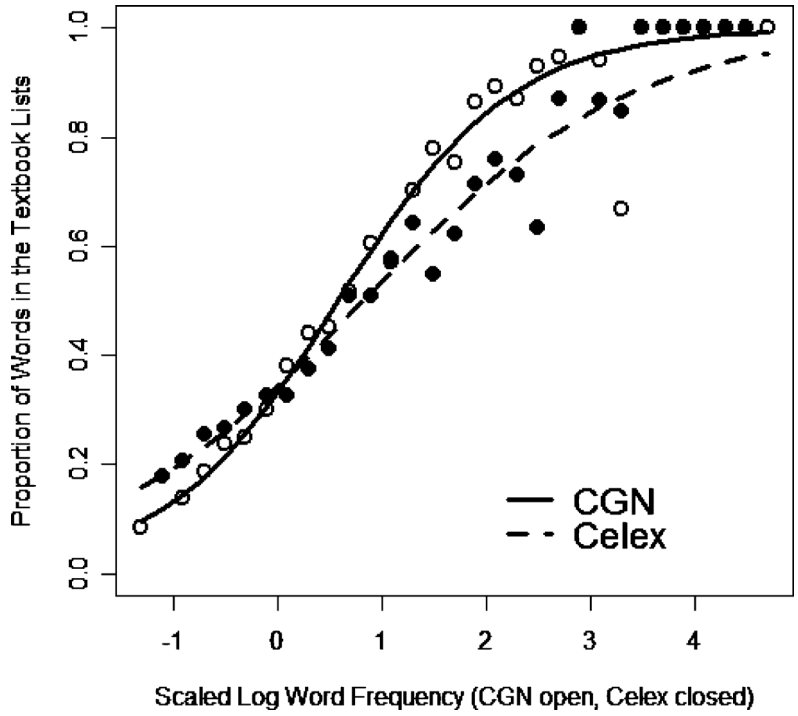

Figure 1. Status as a textbook vocabulary item (0 = no, 1 = yes) based on form frequency. Solid line is the predicted proportion based on the fitted model with CGN as a predictor, dashed line with CELEX; open circles are the empirical proportions for CGN calculated at histogram intervals of the centered and scaled frequency distribution, filled circles are the CELEX proportions.

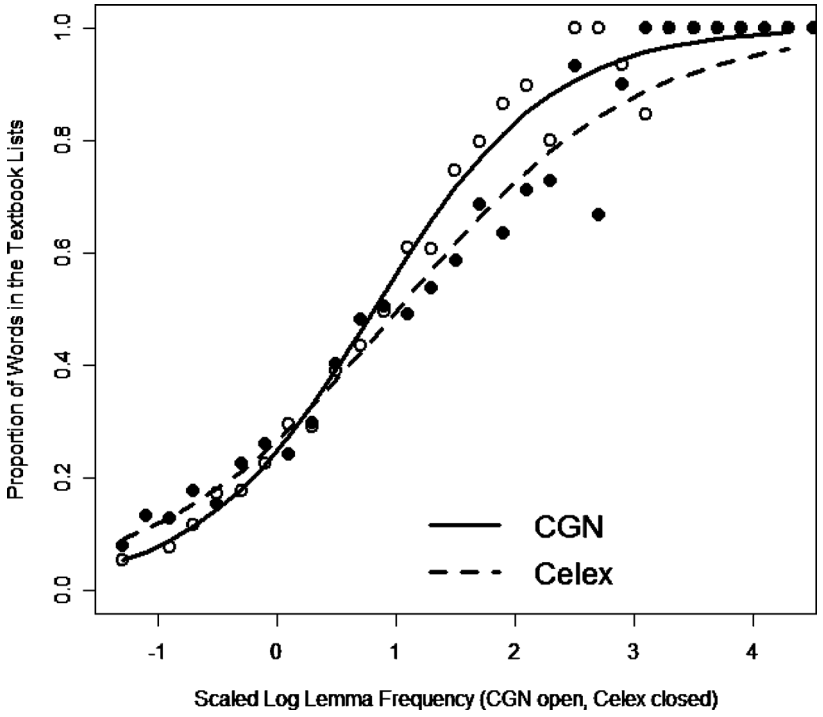

Figure 2. Status as a textbook vocabulary item (0 = no, 1 = yes) based on lemma frequency. Solid line is the predicted proportion based on the fitted model with CGN as a predictor, dashed line with CELEX; open circles are the empirical proportions for CGN calculated at histogram intervals of the centered and scaled frequency distribution, filled circles are the CELEX proportions.

model were Proportion_Textbook = −1.1071 + 1.3528 * Frequency_CGN.

Appendix A contains a list of the 200 noun lemmas with the highest predicted probability of being found in a textbook based on this equation. For example, for the lemma *tijd* "time", the predicted proportion of 0.9833 is obtained by log transforming the CGN frequency (i.e., LOG(1 + 903)) = 6.806829), scaling this value according to the mean (2.224635) and standard deviation (1.195641) of the log CGN frequencies (i.e., (6.806829 – 2.224635)/1.195641 = 3.832417), multiplying this value by the model coefficients (i.e., −1.1071 + 1.3528 * 3.832417 = 4.0774) and substituting this value into the equation for the logistic (i.e., 1/(1 + EXP(−4.0774)) = 0.9833).

In sum, spoken word form frequencies predicted the occurrence of words in learner textbooks better than written word form frequencies, and taking written word form frequencies into account did not improve the prediction based on spoken word forms alone. One interpretation of the results is that while both CELEX and CGN are correlated ($r = 0.765$ for log word form frequencies), there is a difference in variability of the frequency estimates. Apparently, some of the variability of the spoken word form frequencies predicts textbook status, but is not related to the variability in frequency that is shared with the written word form frequencies.

A possible explanation for the better fit of the model that is based on CGN frequency is that the vocabulary lists from the textbooks may include items that are common in colloquial speech, which may not be common in written text for stylistic reasons. Written text is often edited, for example. This could be due to characteristic terms used in speech and not in books, or vice versa. However, not all differences are entirely straightforward or predictable from genre differences: e.g. as seen in Appendix A, the lemma frequency for *soort* "sort", is 1050 in CGN but only 75 in CELEX. Conversely, for *hand* "hand", the frequency is 1028 in CELEX and 342 in CGN. This difference in frequencies would not be expected based on genre alone. This illustrates that different data sources may capture aspects of frequency estimates that are not otherwise easily modeled.

Corpus and textbook predictions of learner usage status

The previous analysis showed that the CGN corpus frequency better predicted whether or not a word appeared in textbook vocabulary lists than CELEX corpus frequency. It can, therefore, be assumed that spoken frequency predicts classroom vocabulary acquisition better than written frequency, given that the appearance of words in teaching materials predicts their likeliness of being learned in tutored SLA. We will now examine to what degree spoken or written corpus frequencies, as well as status as a textbook vocabulary item, predict whether a word is in the active vocabulary of early UNTUTORED learners of Dutch.

It can be expected that for a given sample of words that learners use, a large-corpus frequency estimate would predict whether or not a word appears in the sample because the more likely a word is in a language, the more likely it will be observed, by definition. What is important in the present case is whether or not one frequency estimate is better than another, or whether an ensemble of frequency estimates is better than any single estimate.

In contrast to the situation in classroom learning, which is highly structured according to a syllabus or a textbook, untutored language usage by learners is opportunistic, and structured according to the communicative needs of the learner. It is therefore possible that textbook status will not predict learners' usage, along with, or in addition to, corpus frequency estimates. To investigate this question, we matched the nouns from the textbook list and corpora samples against the nouns appearing in a sample of spoken usage words from the film re-telling task of the Dutch portion of the ESF Second Language Learner corpus (Perdue, 1984, 1993).

Figure 3 shows the fit of the best-fitting model against the empirical proportions of words found in the film re-telling task as a function of form frequency for words found and not found in the textbooks. The best-fitting model included positive terms for both CGN and CELEX word frequency as well as textbook status (AIC = 1298); Proportion_Learner_Usage = −3.9373 + 0.6620 * Textbook + 0.6793 * CGN_Freq + 0.3550 * CELEX_Freq. It should be emphasized that all three data sources were predictive of learner usage. However, in addition, the parameters of this model indicate that the change in proportion of learner usage related to Textbook status and CGN frequency was almost twice as large as the proportional increase related to CELEX frequency. The model with all three data sources was significantly better than models including textbook alone (AIC = 1503; $p < 0.001$), or textbook with CELEX alone (AIC = 1330; $p < 0.001$), or textbook with CGN alone (AIC = 1310; $p < 0.001$). Appendix B contains the forms ranked by the predicted proportion of the best-fitting model.

For a comparison to the previous analysis of frequency as a predictor for textbook status, we also fit models with CGN and CELEX without textbook status as a predictor for learner usage. The model parameters for the single term models (either CGN or CELEX) were of similar sign and magnitude as in the previous analysis. A model that included positive terms for CGN as well as CELEX log frequency was the best fit model (Proportion_Learner_Usage = −3.6489 + 0.8125 * CGN_Freq + 0.3506 * CELEX_Freq; AIC = 1308), compared to CELEX alone (AIC = 1363; $p < 0.001$), or

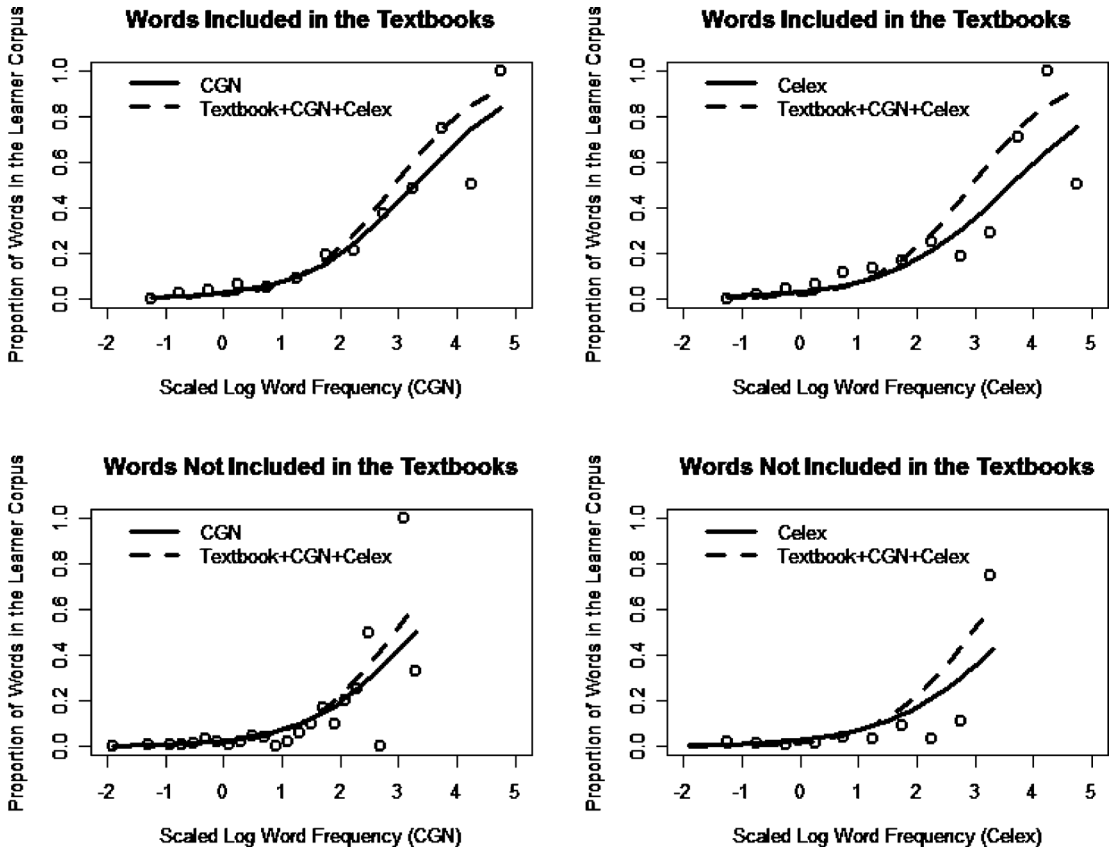

Figure 3. Word usage in the ESF film-retelling task (0 = no, 1 = yes) as function of word frequency and textbook status. Circles are the empirical proportions as calculated at histogram intervals of the centered and scaled frequency distribution.

CGN alone (AIC = 1320; $p < 0.001$). Although there is a difference in the fitted model parameters, it should be emphasized that CELEX frequencies did not greatly improve the model fit in the present analysis, as the difference in AIC between the full model and CGN-alone model is not large (1308 versus 1320).

This analysis revealed that for learner usage during a relatively early period of adult language learning, a combination of predictors is a more effective model of usage than any single predictor. In addition, the CGN word frequency and textbook status were stronger predictors than CELEX word frequencies. Without textbook status in the model, both CGN and CELEX were positive predictors, although again the relationship with CELEX was proportionately smaller.

Discussion

The analyses presented here revealed that both spoken (CGN) and written (CELEX) corpus frequency estimates predicted whether a word would be used as a vocabulary item in textbooks. However, spoken word frequencies predict best what learners are most likely to read in the classroom and including written frequency estimates did not result in a better prediction than spoken frequencies alone.

By contrast, in the analysis of actual (untutored) learner usage the model that predicted best whether speakers used a word in a film re-telling task included both spoken and written corpora and, surprisingly, even the textbook status of words. Given that our first analysis showed that all three sources are highly correlated, this result suggests that what makes combined estimates helpful is the variability in estimated frequency which is not shared among the sources. As a practical consequence, this result shows that no single word frequency count may be best for predicting learner usage or – assuming that usage and comprehension are highly correlated (Laufer et al., 2004) – learner vocabulary knowledge. Instead, a broad sample of many different language use domains should be used, if it is available. In cases where fewer sources are available, the use of textbook status and written frequency estimates alone appear problematic, but their combination was almost as good as spoken frequency

estimates. An interesting extension of the present analysis would be to use frequency of occurrence in a textbook as a predictor variable, if this information could be obtained in a convenient form.

Although the models were fitted to predict occurrence in learner speech about just one specific film topic, it is important to note that they also predict the use of words in other contexts. The words *tijd* "time", *uur* "hour, o'clock", *geval* "case", *leven* "life", *aantal* "number", *jaren* "years", *wereld* "world", *manier* "manner", *vraag* "question", for example, are highly predicted although they are NOT used in the film re-tellings. This is not a flaw of the model. Simply speaking, the models find properties of words (e.g. to have a certain CGN frequency, to occur in textbooks) that make them likely to be used more often during the film re-tellings. Other words that have not been used during the film re-tellings may nonetheless have similar properties, so in a sense they are predicted to be used when talking about an appropriate topic. Of course, even in L2 learners the choice of words used is more strongly determined by what they are talking about than how well they know certain words. It is for the same reason that a broad sample of many different language use domains predicts learner usage best. A broad sample reflects better than any single source what people talk about in different contexts and therefore what is common in a language.

It is quite possible that the first language of an L2 learner could influence the likelihood that low-frequency L2 words are produced. If the L1 and L2 of a learner share cognates which happen to be low frequency, for example, then these words would be produced more often by that learner because of they would be already known (as cognates). A learner of the same L2 but with a different L1 that does not share the cognates would be less likely to produce these words.

Why are words that are more common in a language learned earlier than less common words? Perhaps the simplest explanation might be that commonly encountered words are simply needed more, other factors being equal, in order for effective communication to occur. Experimental research on memory has demonstrated that the so-called NEED PROBABILITY of memory items (Anderson and Schooler, 1991) can be used to account for a variety of effects in research on memory including the shape of practice and forgetting functions, also including those of L2 lexical acquisition (Pavlik and Anderson, 2005). The need probability of an item is simply the probability that some item of information to be learned will be required again in the future. Lexical frequency can be seen as an estimate of the probability that a word will be required again sometime in the future. A productive line of research may be to relate how different aspects of lexical frequency, recency of encounter, and prior exposure over a learner's lifespan may predict learner's acquisition and retention of L2 words.

Several caveats to the approach we have taken are in order. In the present study, we considered whether two relatively large corpora would be useful in predicting learner usage. This might suggest that using even a larger number of data sources would improve predictions even more. However, the cost of obtaining each source has to be weighed against the potential decreasing utility of extending the number of data sources. It is sometimes time-consuming and expensive to obtain additional corpus data (especially with regard to speech data), and if there are already several data sources available, new sources may not add much more predictive ability. Second, if the data that are used are unrepresentative of the target language, it may not help with prediction. With these caveats in mind, although it is unlikely that word frequency alone can serve as a completely adequate predictor, the results presented here suggest that the probability that a word will be encountered in a learner's environment can be most effectively modeled using multiple data sources.

Appendix A. 200 Dutch lemmas with highest prediction of textbook appearance

Lemma	FreqCGN	FreqCELEX	Ntext	Predicted	Lemma	FreqCGN	FreqCELEX	Ntext	Predicted
jaar	1700	1143	3	0.992	politie	126	95	0	0.865
mens	1604	1370	3	0.991	brief	125	200	2	0.864
keer	1548	454	1	0.991	kaart	125	88	3	0.864
soort	1050	75	3	0.986	tent	124	27	1	0.863
zijn	1003	648	3	0.985	student	123	49	2	0.862
tijd	903	1084	3	0.983	grond	122	353	2	0.861
ding	874	371	3	0.983	maandag	122	16	3	0.861
week	834	283	3	0.982	mama	122	43	1	0.861
uur	808	425	3	0.981	straat	122	195	3	0.861
man	729	1190	3	0.979	muziek	121	115	3	0.860
kind	708	961	3	0.978	computer	120	49	2	0.858
punt	576	19	2	0.973	familie	120	134	2	0.858
moment	541	298	3	0.971	telefoon	120	84	3	0.858
stuk	532	46	3	0.970	buurt	118	109	3	0.856
idee	501	211	3	0.968	gebied	118	242	2	0.856
boek	499	387	3	0.968	orde	118	159	0	0.856
huis	495	630	3	0.968	wedstrijd	118	23	1	0.856
paar	488	491	2	0.967	kennis	117	141	3	0.855
geval	457	539	2	0.965	Amsterdam	115	95	0	0.852
jong	440	41	3	0.963	Belgie	114	33	1	0.851
vraag	423	476	3	0.962	gemeente	114	74	3	0.851
probleem	420	340	3	0.961	beeld	113	197	1	0.850
vrouw	404	900	3	0.960	antwoord	112	209	3	0.849
werk	398	571	2	0.959	gezicht	112	504	3	0.849
God	389	298	0	0.958	arm	111	187	1	0.847
school	388	243	3	0.958	reden	109	226	3	0.845
aantal	379	378	3	0.957	fiets	108	48	3	0.843
leerling	373	58	0	0.956	kilometer	108	65	2	0.843
Nederland	347	220	2	0.952	tekst	108	85	3	0.843
moeder	343	596	3	0.952	bedoeling	107	84	0	0.842
hand	342	1028	3	0.952	helft	107	82	3	0.842
kant	339	291	2	0.951	voet	107	225	3	0.842

manier	336	375	3	0.951	vak	105	51	1	0.839
minuut	327	174	3	0.949	gang	104	187	3	0.838
plaats	313	661	3	0.947	programma	104	71	1	0.838
geld	311	280	3	0.947	vorm	104	333	3	0.838
maand	305	230	3	0.945	ruimte	102	150	2	0.835
zin	305	349	2	0.945	voorbeeld	102	229	2	0.835
auto	300	208	3	0.944	kerk	101	205	3	0.833
heer	297	193	1	0.944	donderdag	100	9	3	0.832
vader	282	576	3	0.941	stem	99	307	1	0.830
leven	279	463	3	0.940	einde	98	156	0	0.828
minister	279	111	0	0.940	dinsdag	96	10	3	0.825
avond	278	284	3	0.940	onderwijs	96	162	0	0.825
bal	278	36	2	0.940	moeite	95	148	1	0.823
eind	278	25	1	0.940	mogelijkheid	95	227	1	0.823
meneer	266	205	2	0.937	woensdag	95	9	3	0.823
groep	263	323	3	0.936	Frankrijk	94	84	2	0.822
zaak	252	424	2	0.933	Brussel	93	33	0	0.820
naam	239	420	3	0.929	situatie	93	228	3	0.820
stad	236	323	3	0.928	dorp	92	137	3	0.818
kamer	223	365	3	0.924	baan	91	74	3	0.816
oog	220	820	3	0.923	ogenblik	91	201	1	0.816
verhaal	217	238	2	0.922	tuin	91	119	3	0.816
hoop	216	84	1	0.921	hond	90	168	2	0.815
water	215	364	3	0.921	papa	90	40	1	0.815
feit	211	354	1	0.920	bank	88	114	3	0.811
gulden	210	58	1	0.919	feest	88	60	3	0.811
begin	202	93	2	0.916	persoon	88	195	2	0.811
ouder	201	214	3	0.915	licht	87	276	1	0.809
procent	199	59	2	0.915	rol	87	210	2	0.809
land	196	422	3	0.913	aandacht	86	199	2	0.807
gesprek	190	155	3	0.910	kleed	86	51	2	0.807
hoofd	186	544	3	0.908	niveau	86	105	0	0.807
mevrouw	185	166	3	0.908	periode	86	134	2	0.807

Appendix A. (cont.)

Lemma	FreqCGN	FreqCELEX	Ntext	Predicted	Lemma	FreqCGN	FreqCELEX	Ntext	Predicted
rest	185	115	2	0.908	reis	86	98	3	0.807
klas	180	48	2	0.905	plek	85	90	1	0.805
meisje	180	357	3	0.905	zoon	85	189	3	0.805
vriend	178	284	3	0.904	a	84	47	0	0.803
film	174	106	3	0.902	onderwerp	84	93	2	0.803
weekend	173	17	3	0.901	vriendin	84	72	3	0.803
wereld	167	454	1	0.898	informatie	83	109	3	0.800
bus	166	40	3	0.897	wet	83	187	0	0.800
les	166	32	2	0.897	lijn	82	104	3	0.798
deel	164	388	2	0.896	muur	82	147	3	0.798
deur	162	376	3	0.895	nieuws	82	29	3	0.798
onderzoek	160	204	1	0.893	principe	82	77	1	0.798
foto	155	107	3	0.890	winkel	82	60	3	0.798
voorzitter	155	48	0	0.890	blik	81	188	1	0.796
bedrijf	153	119	3	0.888	eeuw	81	229	2	0.796
jongen	151	360	2	0.887	Frans	81	43	1	0.796
nacht	151	266	3	0.887	ziekenhuis	81	94	3	0.796
vrijdag	151	17	3	0.887	zon	81	46	3	0.796
zaterdag	146	21	3	0.883	dochter	80	120	3	0.794
kans	144	202	0	0.882	indruk	80	155	2	0.794
ure	140	39	0	0.878	rug	80	180	3	0.794
bed	139	300	3	0.877	staat	80	290	0	0.794
trein	138	81	3	0.876	Engels	79	35	1	0.791
verschil	138	175	3	0.876	overheid	79	127	1	0.791
partij	136	170	0	0.875	rekening	79	114	2	0.791
weer	136	16	3	0.875	steen	79	12	1	0.791
tafel	135	247	3	0.874	afspraak	78	53	2	0.789
zondag	131	41	3	0.870	kop	78	135	2	0.789
regering	130	115	0	0.869	papier	78	113	3	0.789
taal	130	156	2	0.869	stof	78	6	0	0.789
richting	129	199	2	0.868	been	77	178	3	0.787
broer	128	128	3	0.867	eten	77	103	2	0.787
gevoel	128	251	3	0.867	raam	77	174	3	0.787
krant	128	117	3	0.867	lichaam	75	292	2	0.782
nummer	128	72	3	0.867	Nederlander	75	35	2	0.782

Appendix B. 200 Dutch word forms with highest prediction of appearance in learner production

Form	CGN	CELEX	NTxt	Film	Predicted	Form	CGN	CELEX	NTxt	Film	Predicted
mensen	1469	934	3	1	0.7131	reden	86	164	3	0	0.2357
tijd	807	958	3	0	0.6802	hart	61	183	2	0	0.2323
jaar	1438	734	3	1	0.6746	vragen	130	134	3	1	0.2309
man	556	876	3	1	0.6414	verband	64	177	1	0	0.2304
zijn	1003	648	3	1	0.6303	straat	102	147	3	1	0.2304
keer	1493	426	1	1	0.5852	blik	76	164	1	0	0.2294
dag	600	607	2	1	0.5834	kennis	108	141	3	0	0.2280
huis	431	541	3	1	0.5391	moeite	95	148	1	0	0.2276
kinderen	491	484	3	1	0.5284	boeken	151	121	3	0	0.2255
paar	486	483	2	1	0.5273	jongens	133	127	2	1	0.2252
moeder	325	555	3	1	0.5233	oog	60	171	3	0	0.2228
hand	215	645	3	1	0.5207	ruimte	100	139	2	0	0.2223
vrouw	258	597	3	1	0.5199	plan	98	137	3	0	0.2195
plaats	264	590	3	1	0.5194	gesprek	147	116	3	0	0.2189
uur	744	380	3	0	0.5147	taal	109	130	2	0	0.2182
werk	380	496	2	1	0.5143	zoon	72	151	3	0	0.2164
vader	271	547	3	1	0.5075	beeld	77	146	1	0	0.2155
ogen	160	649	3	1	0.5004	pijn	71	149	2	1	0.2140
geval	425	411	2	0	0.4882	indruk	78	143	2	0	0.2136
leven	277	443	3	0	0.4708	meter	153	108	2	1	0.2122
dingen	724	298	3	1	0.4686	familie	112	122	2	1	0.2118
hoofd	179	515	3	1	0.4666	vriend	74	143	3	1	0.2111
aantal	371	364	3	0	0.4564	liefde	49	168	1	0	0.2108
jaren	262	406	3	0	0.4511	rest	181	98	2	0	0.2086
kind	196	454	3	1	0.4503	muziek	121	115	3	0	0.2084
wereld	164	444	1	0	0.4336	begin	202	91	2	1	0.2050
manier	316	333	3	0	0.4290	geschiedenis	70	135	2	0	0.2016
vraag	293	342	3	0	0.4284	toekomst	74	131	1	0	0.2007
moment	516	270	3	1	0.4266	zee	64	138	3	1	0.2001
mens	135	436	3	1	0.4166	hoop	216	84	1	0	0.1989
water	213	353	3	1	0.4114	vrienden	80	125	3	0	0.1989
zin	284	312	2	0	0.4099	verschil	116	107	3	0	0.1980
weg	158	386	3	1	0.4062	haar	934	46	3	1	0.1979

Appendix B. (cont.)

Form	CGN	CELEX	NTxt	Film	Predicted	Form	CGN	CELEX	NTxt	Film	Predicted
gezicht	99	448	3	0	0.3998	systeem	60	138	1	0	0.1971
geld	311	276	3	1	0.3949	buurt	118	105	3	0	0.1967
kamer	180	318	3	1	0.3816	wijn	53	141	3	0	0.1940
handen	117	377	3	1	0.3814	periode	81	119	2	0	0.1938
land	196	298	3	0	0.3763	landen	85	114	3	0	0.1912
deur	139	325	3	1	0.3678	dokter	60	130	2	0	0.1904
boek	262	250	3	0	0.3662	aarde	37	153	1	0	0.1873
woorden	152	310	3	0	0.3658	brief	77	114	2	1	0.1868
naam	170	294	3	0	0.3644	kop	71	117	2	1	0.1861
woord	188	282	3	0	0.3642	overheid	71	117	1	0	0.1861
kant	295	235	2	1	0.3638	voeten	54	129	3	0	0.1849
deel	143	312	2	0	0.3628	meisjes	61	120	3	1	0.1823
mannen	145	296	3	0	0.3550	informatie	83	106	3	0	0.1821
grond	117	321	2	1	0.3543	beweging	38	144	1	0	0.1818
Nederland	346	202	2	0	0.3492	dienst	54	125	2	0	0.1815
school	345	198	3	0	0.3458	koffie	72	111	3	1	0.1810
stad	188	249	3	0	0.3435	werkelijkheid	32	152	1	0	0.1804
week	609	155	3	0	0.3430	rekening	75	107	2	0	0.1788
bed	133	284	3	0	0.3425	staat	70	273	0	1	0.1783
vrouwen	128	288	3	0	0.3423	eten	77	103	2	1	0.1759
meneer	266	203	2	1	0.3329	beleid	51	121	1	0	0.1757
gevoel	128	251	3	0	0.3202	benen	42	130	3	0	0.1752
ouders	193	206	3	0	0.3149	film	142	80	3	1	0.1751
idee	468	144	3	1	0.3144	tuin	84	98	3	1	0.1744
groep	180	201	3	0	0.3067	glas	44	124	1	0	0.1722
meisje	119	237	3	1	0.3067	maand	156	74	3	0	0.1710
onderzoek	160	204	1	0	0.3018	avonds	144	76	3	0	0.1704
licht	80	268	1	0	0.3013	persoon	59	108	2	1	0.1699
auto	256	165	3	1	0.2981	afstand	48	117	3	0	0.1697
stem	77	263	1	0	0.2962	mogelijkhede	53	112	1	0	0.1693
probleem	273	158	3	1	0.2954	ervaring	54	111	2	0	0.1691
lichaam	69	264	2	0	0.2902	feite	82	94	1	0	0.1691

wijze	65	268	1	0	0.2889	hond	59	107	2	0	0.1690
vorm	84	242	3	0	0.2889	basis	64	102	2	0	0.1674
zaken	127	201	2	0	0.2857	stoel	45	117	2	1	0.1672
zaak	115	204	2	0	0.2821	broer	77	94	3	0	0.1666
avond	124	195	3	0	0.2798	relatie	51	110	2	0	0.1659
problemen	147	180	3	0	0.2781	leeftijd	64	100	2	0	0.1654
mevrouw	185	164	3	1	0.2779	samenleving	42	118	1	0	0.1653
feit	108	203	1	0	0.2777	tante	67	97	3	0	0.1642
minuten	281	136	3	0	0.2750	raam	46	112	3	1	0.1637
dagen	306	325	0	1	0.2739	prijs	125	75	2	1	0.1634
God	376	298	0	0	0.2733	dorp	64	98	3	0	0.1634
tafel	115	189	3	1	0.2710	wind	52	106	2	0	0.1630
verhaal	165	161	2	1	0.2687	gevallen	32	128	2	1	0.1628
richting	119	183	2	0	0.2683	mogelijkheid	42	115	1	0	0.1627
mond	70	220	2	0	0.2645	ding	150	69	3	1	0.1624
aandacht	86	199	2	0	0.2619	Frankrijk	94	83	2	0	0.1621
dood	57	234	3	1	0.2618	geluk	51	105	2	0	0.1613
nacht	89	191	3	0	0.2581	helft	107	78	3	0	0.1611
maanden	149	155	3	1	0.2576	keuken	74	90	3	0	0.1607
gebied	96	183	2	0	0.2563	dieren	60	97	2	0	0.1599
weken	225	128	3	0	0.2540	hulp	39	115	2	0	0.1599
antwoord	93	177	3	0	0.2500	stuk	478	42	3	1	0.1599
gebruik	64	205	1	0	0.2499	raad	28	130	1	0	0.1592
rol	81	183	2	0	0.2471	spel	58	97	1	0	0.1586
ogenblik	88	177	1	0	0.2471	dochter	64	93	3	1	0.1583
situatie	76	185	3	0	0.2452	artikel	40	112	2	0	0.1583
gang	97	167	3	0	0.2445	taak	35	117	1	0	0.1575
heer	258	111	1	0	0.2422	procent	199	58	2	0	0.1567
eeuw	69	186	2	0	0.2408	leden	35	116	2	0	0.1566
kerk	86	170	3	0	0.2405	middel	25	130	2	0	0.1550
oorlog	68	184	2	0	0.2386	lid	35	114	2	0	0.1550
voorbeeld	77	175	2	0	0.2385	geluid	43	105	1	0	0.1549
rug	78	173	3	0	0.2377	voet	53	96	3	0	0.1542
lucht	69	181	2	0	0.2372	armen	80	81	1	0	0.1536
recht	44	214	1	0	0.2359	Europa	73	84	1	0	0.1536

References

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control*, AU-19, 716–722.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2, 396–408.

Baayen, R. H., Piepenbrock, R. & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Bossers, B. (1996). Woordenschat. In J. H. Hulstijn, R. Stumpel, B. Bossers & C. Van Veen (eds.), *Nederlands als tweede taal in de volwasseneneducatie: Handboek voor docenten*, pp. 167–193. Amsterdam: Meulenhoff Educatief.

Brown, C. (1993). Factors affecting the acquisition of vocabulary: Frequency and saliency of words. In T. Huckin, M. Haynes & J. Coady (eds.), *Second language reading and vocabulary learning*, pp. 263–286. Norwood, NJ: Ablex.

Corpus Gesproken Nederlands. Copyright Nederlandse Taalunie 2004. http://lands.let.kun.nl/cgn/home.htm (accessed 17 October 2007).

Day, R., Omura, C. & Hiramatsu, M. (1991). Incidental EFL vocabulary learning and reading. *Reading in a Foreign Language*, 7, 541–551.

De Kleijn, P. & Nieuwborg, E. (1983). *Basiswoordenboek Nederlands*. Leuven: Wolters.

Donaldson, B. (1996). *Colloquial Dutch: The complete course for beginners*. New York: Routledge.

Dunn, L. M. & Dunn, L. M. (1997). *The Peabody Picture Vocabulary Test – 3rd edition*. Circle Pines, MN: American Guidance Service.

Dupuy, B. & Krashen, S. (1993). Incidental vocabulary acquisition in French as a foreign language. *Applied Language Learning*, 4, 55–63.

Fukkink, R. G., Hulstijn, J. & Simis, A. (2005). Does training of second-language word recognition skills affect reading comprehension? An experimental study. *The Modern Language Journal*, 89, 54–75.

Van Gelderen, A., Schoonen, R., De Glopper, K., Hulstijn, J., Simis, A., Snellings, P. & Stevenson, M. (2004). Linguistic knowledge, processing speed, and metacognitive knowledge in first- and second-language reading comprehension: A componential analysis. *Journal of Educational Psychology*, 96, 19–30.

Hazenberg, S. (1994). Een keur van woorden. Ph.D. dissertation, Vrije Universiteit Amsterdam.

Hulstijn, J. H., Hollander, M. & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal*, 80, 327–339.

van Kampen, H. & Stumpel, R. (2002). *Dutch for self-study/Nederlands voor anderstaligen* (4th edn.). Utrecht: Prisma (Het Spectrum B.V.).

Laufer, B., Elder, C., Hill, K. & Congdon, P. (2004). Size and strength: Do we need both to measure vocabulary knowledge? *Language Testing*, 21, 202–226.

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics*, 26 (1), 32–47.

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (eds.), *LREC-2000 (Second International Conference on Language Resources and Evaluation) Proceedings*, vol. 2, pp. 887–894. Paris: European Language Resources Association.

Pavlik, P. I., Jr. & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29, 559–586.

Perdue, C. (ed.) (1984). *Second language acquisition by adult immigrants: A field manual*. Rowley: Newbury House.

Perdue C. (ed.) (1993). *Adult language acquisition: Cross-linguistic perspectives* (vol. 1: *Field methods*). Cambridge: Cambridge University Press.

Pitts, M., White, H. & Krashen, S. (1989). Acquiring second language vocabulary through reading: A replication of the Clockwork Orange study using second language acquirers. *Reading in a Foreign Language*, 5, 271–275.

R Development Team (2005). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Learning*, 21, 589–619.

Schneider-Broekmans, J. (2000). *Taal vitaal: Nederlands voor beginners*. Amsterdam & Antwerpen: Intertaal.

Sciarone, A. G. (1979). *Woordjes leren in het vreemdetalenonderwijs*. Muiderberg: Coutinho.

Uit den Bogaart, P. C. (ed.) (1975). *Woordfrequenties in geschreven en gesproken Nederlands*. Utrecht: Oosthoek, Scheltema & Holkema.

Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. New York: Springer.

Vermeer, A. (2001). Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics*, 22, 217–234.

Received October 16, 2006
Revision received February 18, 2007
Accepted April 20, 2007