APHASIOLOGY, 1998, VOL. 12, NO. 7/8, 601–617

Beyond frequency: predicting auditory word recognition in normal elderly adults

ANNETTE BAUMGAERTNER and
CONNIE A. TOMPKINS
University of Pittsburgh, Pittsburgh, PA, USA

Abstract

This study examined the contributions of several lexical variables to prediction of the variance in spoken word recognition performance in a sample of 29 normal older adults. Subjects responded to 50 experimental stimulus words varying in frequency, age-of-acquisition, and familiarity, in a speeded auditory lexical decision task. The contributions of familiarity and age of acquisition were examined after accounting for word frequency, the variable most often controlled in such studies. Strong age of acquisition effects were observed after accounting for frequency, whereas familiarity did not contribute to predicting lexical decision reaction times. Clinical and research implications are discussed.

Introduction

The lexical characteristics of language traditionally have been viewed as important factors in understanding language representation and processes. One recent example, relevant to clinical aphasiology, is found in a study by Nickels and Howard (1995) which attempted to relate the difficulties that aphasic patients have with naming particular words to the lexical properties of these words, for example, their frequency, familiarity, and age of acquisition. Of all the variables hypothesized to influence lexical processes, word frequency probably has received the most attention in psycholinguistic studies of word recognition. Word frequency, obtained from counts of occurrences of words in a pre-specified pool of written language samples, has been shown to be a significant factor in various studies of visual and auditory perception and memory for linguistic material. The word frequency effect, denoting significantly faster and more accurate word recognition performance for high-frequency than for low-frequency words, has been demonstrated in almost all word recognition tasks (for a review see Monsell 1991).

The mechanisms underlying the word frequency effect, however, are still subject to controversy (Lively *et al.* 1994). Frequency effects have been suggested to originate in both lexical access processes and in post-access mechanisms (e.g. Balota and Chumbley 1984, Luce *et al.* 1990, Connine *et al.* 1990). And, even in assuming a lexical access locus of the frequency effect, there is no consensus about whether frequency affects the level of activation, the rate of activation, or the

Address correspondence to: Annette Baumgaertner, University of Pittburgh, Department of Communication Science and Disorders, 4033 Forbes Tower, Pittsburgh, PA 15260, USA.

In this study, we asked groups of raters (similar in age and education to the older adults participating in the lexical decision task) to generate estimates of AOA and FAM for a pool of potential stimulus words. In addition to an analysis of the expected effects of FAM and/or AOA, we planned to address the more practical issue of generalizability and replicability of FAM and AOA ratings by comparing those generated for this study with already published ratings by college students (Gilhooly and Logie 1980).

Method

Subjects

Thirty-five adults between the ages of 52 and 74 years participated (mean age 64·1 years, mean education 14·7 years; see table 1 for subject data). All subjects were monolingual native American English speakers, except one subject who spoke British English. All subjects were right-handed, as indicated by verbal response to the six most discriminating items from the Annet (1970) inventory (after Geffen 1982). Right-hand preference was operationally defined as performing all actions with the right hand only. Further, all subjects passed a pure-tone air-conduction hearing screening (35, 35, and 40 dB HL [ANSI 1969] at 500, 1000, and 2000 Hz respectively in the better ear). In addition, all subjects correctly repeated nine one- and two-syllable words containing clusters of high-frequency consonants. (Two potential subjects who had passed the hearing screening were excluded after failing this criterion.) Subjects were questioned to rule out previous major neurological conditions and substance abuse. To be included in the study, subjects had to score at least 27 points on the Mini-Mental State Examination (MMSE, Folstein *et al.* 1975). Three potential subjects did not meet this criterion and were not included in the study. All subjects completed questionnaires addressing social status (Hollingshead Index) and physical and mental health status (SF-36, see table 1 for data and references). Subjects were identified through senior citizen groups, media advertisements, and volunteer departments of local rehabilitation centres.

Experimental task

Auditory lexical decision was chosen as the experimental task for several reasons. The lexical decision task has been used widely as an index of lexical access and has been shown to be sensitive to lexical variables such as frequency and semantic similarity. It is a relatively immediate measure of language processing and permits sensitive reaction time measures in addition to accuracy. Furthermore, it allows manipulation of depth and type of processing by way of the choice of non-word or distractor items. The auditory version of the task was chosen because it was of interest to examine the generality of AOA and FAM effects previously reported for visual lexical decision in the auditory modality. In addition, much of the research in language processing with brain-damaged subjects involves auditory comprehension; the matching of auditorily presented stimuli in terms of their lexical characteristics is a recurrent problem.

The criterion variable was reaction time (RT). To determine the point from which to measure RT, the distribution of word lengths (in ms) in the experimental word set was examined for each of the predictor variables AOA, FAM, and

Table 1. Descriptive data for subject group (n = 29)[a]

Age (years)	
M	63·8
SD	7·9
Range	52–74
Education (years)	
M	14·5
SD	2·6
Range	12–20
Gender	18 female
Mini-Mental State Examination (max. 30)	
M	29·0
SD	0·9
Range	27–30
Peabody Picture Vocabulary Test[b] (max. 175)	
M	169·2
SD	4·4
Range	156–174
Hollingshead Index of Social Status[c] (max. 66)	
M	48·8
SD	11·3
Range	25·0–66·0
Physical Health Status[d]	
M	49·38
SD	9·11
Range	21·66–59·47
Mental Health Status[e]	
M	53·30
SD	9·83
Range	25·60–63·53

[a] Six of the original subjects were excluded; see 'Preliminary analyses' and Appendix 1.
[b] Dunn and Dunn (1981).
[c] Four-factor index of social status (Hollingshead 1975).
[d] SF-36 Physical Component Scale (Ware *et al.* 1994). For general US population, $M = 50$, SD = 10.
[e] SF-36 Mental Component Scale (see reference above).

frequency. Graphs plotting length with each of the predictors showed random distributions of length across the ranges of each variable. We thus concluded that word length could be ruled out as a confounding variable and decided to measure RT from word onset. Mean word length was 537·2 ms (range 420–715 ms, SD 71·4 ms).

Experimental stimuli

The creation of experimental stimuli proceeded in two steps. First, an initial pool was screened to select a set of potential word candidates. Specific screening criteria are discussed below. In the second step, FAM and AOA ratings were obtained and the stimulus set size was reduced to 50 words. In obtaining the final set of 50 words, care was taken to maintain a sample sufficiently heterogenous for the predictor variables word frequency, AOA, and FAM.

Selection of word candidates

An existing corpus of 1944 words (Gilhooly and Logie 1980) was used to select an initial pool of about 200 nouns. Only two-syllable words, stressed on the first syllable, were selected. Proper names and words that were ambiguous or homophonic were not included. Also, care was taken not to include words that might be completely unfamiliar (e.g. *unction*). Single words representing more than one part of speech (verbs as well as nouns) were included, as long as their semantic content did not change (e.g. *visit* was included, but not *balance*).

Several criteria were then applied to reduce the initial word pool. Only words consisting of four to six phonemes, following standard international phonetic convention, were included. This resulted in a sample of 66 words with five or six phonemes and 34 words with four phonemes.

These words were then examined for neighbourhood (N) size and N frequency. Conventionally, a 'neighbour' is defined as any word resulting by changing one letter of the stimulus word while preserving letter position (Andrews 1989, Coltheart *et al.* 1977, Sears *et al.* 1995).[1] As an example, *truffle*, *tribal*, and *treble* are all (phonological) neighbours of the word *trouble*, following standard English pronunciation (Gimson and Cruttenden 1994). N size is operationalized as the sum of neighbours of a word and N frequency as the frequency of a word's neighbour relative to its own frequency (Grainger 1990, Sears *et al.* 1995). There are conflicting findings with respect to the effects of N size and N frequency on speed and accuracy of lexical decision (see Forster and Shen 1996, for a review). However, regardless of the ongoing controversy about the nature and locus of the effects, N size and frequency appear to influence response times in lexical decision tasks, particularly so when the experimental stimuli vary in word frequency (Andrews 1989, Grainger 1990, Sears *et al.* 1995). Most relevant for our purposes is the finding that N size does not influence RT in a lexical decision task when the N size is small and when none of the neighbours is of a higher frequency than the target (Sears *et al.* 1995). In this study, then, words had to have a N size of less than three to be included, with no neighbours higher in frequency than the respective experimental item. After applying the criteria for N size and N frequency, 37 items were excluded, resulting in 63 potential stimulus words.

As a measure of written word frequency, Carroll *et al.*'s (1971) U-metric parameter, a measure of estimated frequency per one million words, was chosen for several reasons. First, it has been described as the broadest and most current compilation of frequency of occurrence of English terms in natural language (Hayes 1988, Lovelace 1988). Further, it is based on a much larger sample than the

[1] As a cautionary remark, it should be noted that most of the research in this area pertains to the visual modality and that there may be problems with applying similar operationalizations to auditorily presented stimuli. First, in contrast to identifying neighbours by substituting graphemes of a target word, phoneme boundaries are less clearly defined because of coarticulation. Second, because of the sequential nature of spoken words, neighbourhood size might be operationalized alternatively as the sum of words with the same initial phonological structure as the particular word segment being processed. Thus, the 'neighbourhood' of a spoken word is not a static property (as for a written word) but an attribute that changes as more and more of the word is being heard.

Table 2. AOA and FAM ratings (based on 15 subjects each) and U-metric value for 42 experimental words[a]

Stimulus	Mean AOA	Range	SD	Mean FAM	Range	SD	U-metric
algae	6·47	5–7	0·83	2·87	1–6	1·64	6
angel	2·67	1–5	1·11	5·53	2–7	1·60	5
ankle	2·43[b]	1–5	1·09	6·20	4–7	1·01	6
barrel	3·60	2–6	1·24	5·07[b]	3–7	1·49	14
bottom	2·21[b]	1–3	0·80	6·29[b]	4–7	1·14	155
brother	2·00	1–3	0·76	6·33	4–7	1·05	104
burden	5·40	3–7	1·12	4·87	3–7	1·19	6
comrade	5·60	4–7	1·12	5·64[b]	4–7	1·08	2
elbow	2·20	1–4	1·01	6·33	4–7	0·98	8
future	5·33	3–7	1·29	6·21[b]	4–7	0·97	61
harness	4·60	3–6	1·12	3·53	1–6	1·96	12
knowledge	5·13	3–7	1·06	6·40	4–7	1·06	75
level	4·40	3–7	1·40	5·80	4–7	1·26	91
meadow	3·29[b]	2–4	0·61	5·13	1–7	1·88	16
merit	6·07	5–7	0·83	5·20	3–7	1·32	2
minute	2·79[b]	1–4	0·80	5·40	1–7	2·06	116
music	2·29[b]	1–4	0·73	6·07	4–7	1·28	151
ocean	3·47	1–7	1·60	6·33	3–7	1·23	134
peasant	5·87	4–7	1·06	4·33	1–7	2·06	6
pigeon	3·07	2–5	0·96	6·07[b]	3–7	1·44	6
plaza	6·64[b]	6–7	0·50	4·53	2–6	1·46	1
pleasure	4·53	3–6	1·13	6·13	4–7	1·06	29
poem	3·14[b]	1–5	1·10	5·73	3–7	1·44	60
portion	5·27	3–7	1·28	6·20	5–7	0·68	19
prairie	5·67	3–7	1·18	4·73	1–7	1·91	22
process	5·73	4–7	1·03	5·40	3–7	1·24	77
purpose	4·87	2–7	1·51	6·13	4–7	0·99	64
refuse	6·29[b]	5–7	0·91	4·27	1–7	1·62	9
salad	3·67	2–7	1·40	6·47	4–7	0·92	3
savage	5·27	3–7	1·33	3·60	1–6	1·59	9
shadow	3·33	1–6	1·29	5·40	1–7	1·92	37
shepherd	3·27	2–5	1·10	4·07	1–7	1·83	12
signal	4·47	2–6	1·25	5·47	2–7	1·60	47
temper	4·13	2–7	1·41	6·29[b]	4–7	0·99	9
timber	4·60	3–6	0·99	5·86[b]	4–7	1·10	16
value	5·00	3–7	1·13	6·47	5–7	0·74	80
venom	5·80	4–7	1·01	4·07	1–7	1·98	1
vigil	5·80	3–7	1·42	3·93	1–7	1·79	1
village	4·53	3–7	1·41	5·87	4–7	1·06	117
visit	2·80	1–4	1·01	6·20	3–7	1·26	81
whisper	2·53	1–4	1·06	5·47	2–7	1·55	11
woman	2·79[b]	1–5	1·12	6·73	5–7	0·59	126

AOA: $M = 4{\cdot}26$; range = 2·00–6·64; SD = 1·37. FAM: $M = 5{\cdot}44$; range = 2·87–6·73; SD = 0·96.

[a] Eight of the original 50 words were excluded; see 'Preliminary analyses' and Appendix 1.

[b] Means based on 14 ratings after excluding one outlier; see 'Stimulus rating procedures'.

ratings distribution, whereas all FAM outliers occurred in the lower tail ('very unfamiliar') of the FAM ratings distribution.

Following the computations of means for AOA and FAM, the initial sample of 63 words was reduced to a set of 50 words. This was done to arrive at a sample size representing a manageable amount of testing material for a session of about 60–70

Kucera and Francis (1967) frequency count, which is particularly biased towards low-frequency words. A larger U-metric value indicates a higher frequency of occurrence in the word count. The 63 words ranged from 1 to 155 on the U-metric scale, or 99·3 % of the frequency range in the Carroll *et al.* corpus. This proportion was considered sufficiently large to represent meaningful differences in terms of frequency among the experimental items. All words with a count of < 1 in the Carroll *et al.* corpus were set to 1.

Stimulus rating procedures

The 63 stimulus words were submitted to AOA and FAM ratings by judges chosen to match the anticipated experimental subject group in age, gender, and education. Two groups each of 15 older adults participated in rating FAM and AOA. The FAM rating group did not differ from the AOA rating group in age or education: mean age was 68·7 years (67·3 for AOA) and mean education was 14·5 years (14·7 for AOA). One person participated in both groups.

Typically, a seven-point scale, ascending in 2 year increments from age 0–2 to 13 years and older, is used to quantify AOA judgments (Brown and Watson 1987, Gilhooly and Logie 1980). This scale has also been used in recent studies examining AOA effects (Morrison and Ellis 1995, Morrison *et al.* 1992). Piloting of the rating procedure showed that rating a particular age *interval* by means of discrete marks could be confusing to subjects. Accordingly, seven boxes were used to represent the seven age intervals. Subjects were asked to estimate the age interval in which they learned the stimulus words.

FAM ratings also have been done conventionally with seven-point rating scales (Gilhooly and Logie 1980, McCloskey 1980, Nusbaum *et al.* 1984). As in the AOA rating scale, boxes instead of marks were used. The end-points of the rating scale were marked 'very unfamiliar' and 'highly familiar'. We operationalized 'familiarity' as a measure of the extent and type of experience subjects felt they had with each word. Subjects were instructed to rate as 'familiar' words that are common, which they hear and use often, and with which they felt they had more experience.

The raters for AOA were contacted by telephone. The procedure was explained and the rating form was mailed to them. Rating of FAM was done in the second author's laboratory. The rating was self-paced and done independently. Each rating list included explicit instructions and four pre-rated examples.

Next, mean AOA and FAM were computed for each item (see table 2). The mean was chosen despite the ordinal character of the rating scores, for two reasons. First, using the median would have resulted in many tied ranks that would not have reflected the differences in ratings, particularly in the tails of the scoring distributions for each word. In addition, it has been shown that when approximate equidistance can be assumed between ranks (as is assumed for the AOA and FAM ratings), ordinal variables may be treated as interval variables in correlational analyses (Labovitz 1970). Outlying ratings were excluded from the mean calculations for 17 % of the stimuli rated for AOA and 16 % of those rated for FAM (see table 2). An outlier was defined as any single rating that deviated by at least two points from any other individual rating. For example, whereas 14 subjects gave the word 'music' AOA ratings that ranged from one to four, one person rated it as six, and this value was considered an outlier. AOA outliers occurred in both tails of the

Table 2. AOA and FAM ratings (based on 15 subjects each) and U-metric value for 42 experimental words[a]

Stimulus	Mean AOA	Range	SD	Mean FAM	Range	SD	U-metric
algae	6·47	5–7	0·83	2·87	1–6	1·64	6
angel	2·67	1–5	1·11	5·53	2–7	1·60	5
ankle	2·43[b]	1–5	1·09	6·20	4–7	1·01	6
barrel	3·60	2–6	1·24	5·07[b]	3–7	1·49	14
bottom	2·21[b]	1–3	0·80	6·29[b]	4–7	1·14	155
brother	2·00	1–3	0·76	6·33	4–7	1·05	104
burden	5·40	3–7	1·12	4·87	3–7	1·19	6
comrade	5·60	4–7	1·12	5·64[b]	4–7	1·08	2
elbow	2·20	1–4	1·01	6·33	4–7	0·98	8
future	5·33	3–7	1·29	6·21[b]	4–7	0·97	61
harness	4·60	3–6	1·12	3·53	1–6	1·96	12
knowledge	5·13	3–7	1·06	6·40	4–7	1·06	75
level	4·40	3–7	1·40	5·80	4–7	1·26	91
meadow	3·29[b]	2–4	0·61	5·13	1–7	1·88	16
merit	6·07	5–7	0·83	5·20	3–7	1·32	2
minute	2·79[b]	1–4	0·80	5·40	1–7	2·06	116
music	2·29[b]	1–4	0·73	6·07	4–7	1·28	151
ocean	3·47	1–7	1·60	6·33	3–7	1·23	134
peasant	5·87	4–7	1·06	4·33	1–7	2·06	6
pigeon	3·07	2–5	0·96	6·07[b]	3–7	1·44	6
plaza	6·64[b]	6–7	0·50	4·53	2–6	1·46	1
pleasure	4·53	3–6	1·13	6·13	4–7	1·06	29
poem	3·14[b]	1–5	1·10	5·73	3–7	1·44	60
portion	5·27	3–7	1·28	6·20	5–7	0·68	19
prairie	5·67	3–7	1·18	4·73	1–7	1·91	22
process	5·73	4–7	1·03	5·40	3–7	1·24	77
purpose	4·87	2–7	1·51	6·13	4–7	0·99	64
refuse	6·29[b]	5–7	0·91	4·27	1–7	1·62	9
salad	3·67	2–7	1·40	6·47	4–7	0·92	3
savage	5·27	3–7	1·33	3·60	1–6	1·59	9
shadow	3·33	1–6	1·29	5·40	1–7	1·92	37
shepherd	3·27	2–5	1·10	4·07	1–7	1·83	12
signal	4·47	2–6	1·25	5·47	2–7	1·60	47
temper	4·13	2–7	1·41	6·29[b]	4–7	0·99	9
timber	4·60	3–6	0·99	5·86[b]	4–7	1·10	16
value	5·00	3–7	1·13	6·47	5–7	0·74	80
venom	5·80	4–7	1·01	4·07	1–7	1·98	1
vigil	5·80	3–7	1·42	3·93	1–7	1·79	1
village	4·53	3–7	1·41	5·87	4–7	1·06	117
visit	2·80	1–4	1·01	6·20	3–7	1·26	81
whisper	2·53	1–4	1·06	5·47	2–7	1·55	11
woman	2·79[b]	1–5	1·12	6·73	5–7	0·59	126

AOA: M = 4·26; range = 2·00–6·64; SD = 1·37. FAM: M = 5·44; range = 2·87–6·73; SD = 0·96.

[a] Eight of the original 50 words were excluded; see 'Preliminary analyses' and Appendix 1.

[b] Means based on 14 ratings after excluding one outlier; see 'Stimulus rating procedures'.

ratings distribution, whereas all FAM outliers occurred in the lower tail ('very unfamiliar') of the FAM ratings distribution.

Following the computations of means for AOA and FAM, the initial sample of 63 words was reduced to a set of 50 words. This was done to arrive at a sample size representing a manageable amount of testing material for a session of about 60–70 minutes, while being sufficiently large to examine the research question with adequate power. In reducing the set of experimental words, care was taken to obtain maximum spread and approximately normal distributions for each of the predictor variables.

Finally, to evaluate the reliability of the FAM and AOA ratings, the rating task was repeated after 1 year, with subsets of the original raters (for AOA, $N = 7$; for FAM, $N = 5$). Kendall tau values for the final set of experimental items (42 words; see Appendix 1) showed acceptable correspondence only for AOA (0·76; for FAM: 0·61). Paired t-tests indicated no significant difference in ratings over time, for either variable (AOA: t (42) = 0·48, $p = 0·635$; FAM: t (42) = −0·53, $p = 0·601$).

Construction of fillers and non-words

The set of experimental items was complemented with filler and non-word stimuli. The fillers consisted of one-syllable real words. Non-words were one- and two-syllable phoneme strings. They were created by changing the last one or two phonemes of the original words, varying at least two features (e.g. place and manner of articulation). The deviating phoneme(s) always occurred at the end of the phoneme string to induce subjects to postpone their lexical decision until they had heard the entire stimulus. This was done to render the non-words relatively word-like, thus inducing a more 'effortful' way of processing (as opposed to a rather automatic response) to maximize the probability of finding significant effects in the predictor variables (see Forster and Shen 1996, Shoben 1982, for discussions of control in visual lexical decision tasks). In addition, this provided a means to control for different acoustic durations of the words. Making subjects wait until they heard the entire stimulus also balanced for possible effects of different recognition points of the words, or the point at which, depending on their phonologically similar neighbours, words become identifiable (Grosjean 1980, Marslen-Wilson 1987).

Preparation of materials

The final experimental stimulus set of 50 two-syllable words was complemented with 30 one-syllable filler words, along with 60 non-words. The ratio of two- to one-syllable items was kept the same for words and non-words: about 60% of the non-words consisted of two-syllable strings. The total of 140 items was divided into four blocks of 23 and two blocks of 24 items. Experimental items, fillers, and non-word items were pseudorandomly assigned to the blocks. Care was taken to avoid phonological or semantic similarity and coincidence of the same initial phoneme between neighbouring stimuli. No more than three words or non-words and no more than three two-syllable stimuli (words or non-words) occurred in succession. The first two and at least the final stimulus in each block consisted of either non-words or filler items.

Each stimulus was recorded at a normal rate by an experienced native American English speaker (the second author) in a sound-treated booth, using a professional-quality audio-cassette recorder (Marantz PMD420) and high-quality tapes. The stimuli were then transferred to a 486 desktop computer and digitized using mono recording with a sampling rate of 22·05 kHz with 16 bit resolution. Stimuli were computer-edited using *Creative Wave Studio* (Soundblaster 16, Creative Technology Ltd, IBM, Singapore, 1994) software, to insert an alerting trial number and a silent interval between stimulus number and experimental word. A response time initiation pulse was placed at the onset of each word, on a channel inaudible to the subjects. Each trial consisted of an alerting trial number, an interval of 600 ms, and the experimental stimulus. Trials were arranged into blocks, and transferred to a Dell Latitude LX 4100T lap-top computer.

Experimental procedures

Subjects were tested individually for about 60–70 minutes, either in a quiet room at their house, or in the second author's laboratory. Testing began with hearing screenings (audiometric screening and word repetition); if subjects passed criterion, informed consent was obtained. Testing followed a predetermined experimental protocol. The order of presentation of the blocks was randomly assigned for each subject. The blocks were presented in sets of two and interspersed with the descriptive tasks shown in table 1.

Experimental items were played off the lap-top computer, routed to a headphone amplifier (Edcor HA 400C), and presented binaurally through high-quality headphones (Fostex T20) at a comfortable loudness level. Subjects were provided with four live-voice and at least six taped practice items for the experimental task. They were encouraged to respond as fast and accurately as possible. Subjects responded to the experimental items by pressing one of two buttons labelled 'yes' and 'no' on a manual response box. The experimenter initiated each stimulus by pushing a button on the lap-top, keeping inter-stimulus intervals as constant as possible.

Results

Preliminary analyses

Six of the 35 subjects tested and eight of the 50 experimental words were excluded from the final analyses because of errors and equipment failure (see Appendix 1 for details on subject and item exclusion criteria). Demographic data showed no difference between the initial and final subject groups. The final analyses thus included only those experimental items that were correctly recognized by all of the subjects in the study and only the data from subjects who responded correctly to all 42 items. Because of the difficulty of unambiguously identifying RT outliers, and a possible loss of data, it was decided to use the median RT ($N = 29$) for each of the experimental items as the criterion variable (see Appendix 2 for raw RT data).

Before running the main analyses, subjects' responses were inspected for individual errors to check for any extraneous sources of variability in the data. Individual errors on non-words ranged from 0 % to 6·7 % (group mean 2 %, for 29 subjects). There was no relationship between level of education and non-word error frequency. Furthermore, there was no relationship between median RT

Table 3. Pearson product–moment correlations among predictor variables (word frequency, age-of-acquisition, and familiarity) and with median response time (median RT, $n = 29$)

	AOA	FAM	U-metric
FAM	−0·53		
U-metric	−0·41	0·51	
Median RT	0·66	−0·42	−0·46

FAM = familiarity, AOA = age of acquisition, U-metric = word frequency.

Table 4. Hierarchical multiple regression summary tables, outcome variable = median RT, based on 29 subjects ($n = 42$ words)

	R^2	R^2 change	Significance
Model 1			
Step 1 (U-metric)	0·212	0·212	0·002
Step 2 (FAM)	0·260	0·048	0·120
Step 3 (AOA)	0·474	0·214	0·000
Model 2			
Step 1 (U-metric)	0·212	0·212	0·002
Step 2 (AOA)	0·474	0·262	0·000
Step 3 (FAM)	0·474	0·000	0·933

FAM = familiarity, AOA = age of acquisition, U-metric = word frequency.

($N = 29$) and age, education, vocabulary knowledge, physical or mental health, or social status (all $r < 0{\cdot}28$, $p > 0{\cdot}10$).

In preparation for multiple regression analyses, univariate correlations between predictor variables and between predictors and the outcome variable, median RT, were run first. In evaluating these and the following results, it may be helpful to keep in mind that the unit of analysis was the stimulus word. Thus, the variability between words, not between subjects, was of interest. AOA showed the highest correlation with median RT, followed by frequency and FAM. The intercorrelations among the predictor variables were moderate (see table 3).

Primary analyses

The first research question asked how much written word frequency contributed to the variance in RT for auditory lexical decision. As shown in table 3, the correlation of frequency with median RT was −0·46. Thus, frequency explained 21 % of the total variance in RT ($p < 0{\cdot}01$).

The second research question asked whether AOA and/or FAM added significantly to explaining the variance in RT, once frequency was controlled. To examine this question, several hierarchical multiple regression analyses were performed. Median RT was always regressed first on frequency (U-metric). The other variables were added to the regression equations in the order specified in table 4. As evident from the table, AOA made a significant contribution to predicting RT, even after frequency and FAM were accounted for. In contrast to AOA, FAM did not contribute to predicting RT in either model.

Proceeding from these results, we addressed the validity and generalizability of AOA ratings. The ratings generated for this study were correlated with existing ratings done by college students (Gilhooly and Logie 1980). A Spearman rank order coefficient of 0·94 was found for those 44 words for which ratings of both groups were available. An additional multiple regression analysis, replacing the older adults' AOA ratings with the ratings from college students, replicated the main finding for AOA.

Discussion

The first research question in this study addressed the contribution of frequency to explaining variance in auditory lexical decision RTs. Frequency, as the first variable entered in a multiple regression analysis, explained a significant part of the variance. This result was expected, given previous findings reporting frequency effects in visual and auditory word recognition. The primary focus of this study, however, was to examine whether FAM and/or AOA would show an effect after accounting for frequency.

AOA emerged as a strong predictor of performance in auditory lexical decision, even when the effects of frequency and FAM were taken into account. This corroborates previous findings in visual lexical decision, particularly Morrison and Ellis's (1995) suggestion of partially independent influences of word frequency and AOA. Furthermore, our finding is consistent with Forster's (1992) modified serial search model which predicts significant effects of both AOA and frequency. Our results, based on neurologically normal older adults, are paralleled by studies of word recognition in aphasia. Partially independent effects of AOA were found in studies investigating the influence of several lexical variables on naming and reading in aphasic subjects. For example, Nickels and Howard (1995) found that aphasic patients' naming was significantly affected by AOA after accounting for frequency and FAM. Hirsh and Ellis (1994), in a single case study of an aphasic patient, observed significant effects only of AOA on spoken and written naming.

The high correlation between AOA ratings acquired for this study and those already published provides evidence for the validity and generalizability of available AOA data and suggests that aphasiologists may not need to generate individualized ratings for clinical and research applications. More research is needed to assess this possibility.

In contrast to AOA, FAM did not contribute to predicting RT when it was added to the regression model after frequency. Several factors may have led to this finding. Statistically, FAM showed a higher collinearity with frequency than did AOA (0·51 as opposed to −0·41). Frequency, thus, could have taken more of the common variance with it, being the first variable to be entered. In addition, the test–retest reliability for FAM was low, possibly reflecting the difficulty to operationalize FAM. This is consistent with previous reports of diverging operationalizations of FAM and inconsistent rating instructions across studies (Nickels and Howard 1995).

The findings of this study have several implications for evaluating the performance of patients with neurogenic communication disorders on psycholinguistic tasks. For example, the results suggest that evaluating stimuli in terms of rated AOA may be at least as important as assessing them for frequency based on published word counts derived from printed material. Controlling for AOA may be particularly important when tasks involve spoken word recognition. More generally, uncontrolled AOA effects could confound interpretations in a variety of experimental studies. That is, performance differences attributed to contrasting item classes (e.g. object vs. action naming; metaphoric vs. literal interpretation) may be due in part to AOA, or other lexical characteristics.

Future theoretical work will help to delineate the conceptual divergence and overlap among each of the lexical variables considered in this study, as well as the mechanisms responsible for their effects. At the same time, more empirical investigations of AOA appear to be warranted. For instance, it may also be important to examine how well AOA predicts patients' performance on existing clinical assessment tools, similar to Brookshire and Nicholas's (1995) recent investigation of frequency and FAM effects on Boston Naming Test performance. Furthermore, selecting stimuli in terms of rated AOA might be recommended for some treatment approaches; this may be particularly relevant to the treatment of aphasic comprehension disorders in cases where a systematic gradual increase in stimulus difficulty is desired. Finally, future investigation could evaluate the potential of AOA in the treatment context.

References

American National Standards Institute 1969, *Specification for Audiometers, ANSI S3.6–1969* (New York: ANSI).

Andrews, S. 1989, Frequency and neighborhood effects on lexical access: activation or search? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **15**, 802–814.

Annet, M. A. 1970, A classification of hand preference by association analysis. *British Journal of Psychology*, **61**, 303–321.

Balota, D. A. and Chumbley, J. I. 1984, Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340–357.

Brookshire, R. H. and Nicholas, L. E. 1995, Relationship of word frequency in printed materials and judgments of word frequency in daily life to *Boston Naming Test* performance of aphasic adults. *Clinical Aphasiology*, **23**, 107–119.

Brown, G. D. A. and Watson, F. L. 1987, First in, first out: word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition*, **15**, 208–216.

Carroll, J. B., Davies, P. and Richman, B. 1971, *The American Heritage Word Frequency Book* (Houghton Mifflin, Boston).

Coltheart, M., Davelaar, E., Jonasson, J. T. and Besner, D. 1977, Access to the internal lexicon. In S. Domic (Ed.) *Attention and Performance VI* (Erlbaum, Hillsdale, NJ), pp. 535–555.

Connine, C. M., Mullenix, J., Shernoff, E. and Yelen, J. 1990, Word familiarity and frequency in visual and auditory word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **6**, 1084–1096.

Dunn, L. M. and Dunn, L. M. 1981, *Peabody Picture Vocabulary Test – Revised* (Circle Pines, MN: American Guidance Service).

Folstein, M. F., Folstein, S. E. and McHugh, P. R. 1975, Mini mental state. A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, **12**, 189–198.

Forster, K. I. 1992, Memory-addressing mechanisms and lexical access. In R. Frost and L. Katz (Eds) *Orthography, Phonology, Morphology, and Meaning* (Amsterdam: Elsevier Science), pp. 413–434.

Forster, K. I. and Shen, D. 1996, No enemies in the neighborhood: absence of inhibitory neighborhood effects in lexical decision and semantic categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **22**, 696–713.

Geffen, G. 1982, *Dichotic Monitoring Measurement of Hemispheric Specialization for Speech Perception: Manual* (Bedford Park, Australia: The Flinders University of South Australia).

Gernsbacher, M. A. 1984, Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, **113**, 256–281.

Gilhooly, K. J. and Gilhooly, M. L. 1980, The validity of age of acquisition ratings. *British Journal of Psychology*, **71**, 105–110.

Gilhooly, K. J. and Logie, R. H. 1980, Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1944 words. *Behavior Research Methods & Instrumentation*, **12**, 395–427.

Gilhooly, K. J. and Watson, F. L. 1981, Word age-of-acquisition effects: a review. *Current Psychological Reviews*, **1**, 269–286.

Gimson, A. C. and Cruttenden, A. 1994, *Gimson's Pronunciation of English* (London: Edward Arnold).

Grainger, J. 1990, Word frequency and neighborhood frequency effects in lexical decision and naming. *Journal of Memory and Language*, **29**, 228–244.

Grosjean, F. 1980, Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **28**, 267–283.

Hayes, D. P. 1988, Speaking and writing: distinct patterns of word choice. *Journal of Memory and Language*, **27**, 572–585.

Hirsh, K. W. and Ellis, A. W. 1994, Age of acquisition and lexical processing in aphasia: a case study. *Cognitive Neuropsychology*, **11**, 435–458.

Hollingshead, A. B. 1975, Four-Factor Index of Social Status. Unpublished manuscript (Yale University, Department of Sociology, New Haven).

Kucera, H. and Francis, W. N. 1967, *Computational Analysis of Present-Day American English* (Providence, Rhode Island: Brown University Press).

Labovitz, S. 1970, The assignment of numbers to rank order categories. *American Sociological Review*, **35**, 515–524.

Lively, S. E., Pisoni, D. B. and Goldinger, S. D. 1994, Spoken word recognition: research and theory. In Morton Ann Gernsbacher (Ed.) *Handbook of Psycholinguistics* (San Diego: Academic Press), pp. 265–301.

Lovelace, E. A. 1988, On using norms for low-frequency words. *Bulletin of the Psychonomic Society*, **26**, 410–412.

Luce, P. A., Pisoni, D. B. and Goldinger, S. D. 1990, Similarity neighborhoods of spoken words. In Gerry T. M. Altmann (Ed.) *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives* (Cambridge, MA: MIT Press), pp. 122–147.

Marslen-Wilson, W. 1987, Functional parallelism in spoken word-recognition. *Cognition*, **25**, 71–102.

McCloskey, M. 1980, The stimulus familiarity problem in semantic memory research. *Journal of Verbal Learning and Verbal Behavior*, **19**, 485–502.

Metsala, J. L. 1997, An examination of word frequency and neighborhood density in the development of spoken-word recognition. *Memory & Cognition*, **25**, 47–56.

Monsell, S. 1991, The nature and locus of word frequency effects in reading. In D. Besner and G. W. Humphreys (Eds) *Basic Processes in Reading: Visual Word Recognition* (Hillsdale, NJ: Lawrence Erlbaum), pp. 148–197.

Morrison, C. M. and Ellis, A. W. 1995, Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**, 116–133.

Morrison, C. M., Ellis, A. W. and Quinlan, P. T. 1992, Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory and Cognition*, **20**, 705–714.

NICKELS, L. and HOWARD, D. 1995, Aphasic naming: what matters? *Neuropsychologia*, **33**, 1281–1301.

NUSBAUM, H. C., PISONI, D. B. and DAVIS, C. K. 1984, *Sizing up the Hoosier Mental Lexicon: Measuring the Familiarity of 20 000 words.* Research on Speech Perception, Progress Report No 10 (Speech Research Laboratory, Indiana University, Bloomington, Indiana).

SEARS, C. R., HINO, Y. and LUPKER, S. J. 1995, Neighborhood size and neighborhood frequency effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **21**, 876–900.

SHOBEN, E. 1982, Semantic and lexical decisions. In C. Richard Puff (Ed.) *Handbook of Research Methods in Human Memory and Cognition* (New York: Academic Press), pp. 287–314.

WALLEY, A. C. and METSALA, J. L. 1990, The growth of lexical constraints on spoken word recognition. *Perception and Psychophysics*, **47**, 267–280.

WARE, J. E., KOSINSKI, M. and KELLER, S. D. 1994, *SF-36 Physical and Mental Health Summary Scales: A User's Manual* (Boston: The Health Institute).

Appendix 1

Two subjects were excluded because of equipment failure. Four subjects made errors on experimental items (one error per subject) which ranked in the end-points of the FAM, AOA, or frequency distributions (that is, among the lowest three ranks for FAM and U-metric, among the highest three ranks for AOA). It should be noted that these items were recognizable to the rest of the subject group. No more than one of the other 29 subjects, respectively, produced an RT outlier on those items. Level of education did not seem to have an influence on the occurrence of errors, since it ranged from 12 to 18 (mean 15·2 years) in the subjects who made an error. Because the goal was to maintain the largest spread possible for FAM, AOA, and frequency (see 'Stimulus rating procedures'), it was critical to keep these items, and the four subjects were thus excluded from the analysis.

Six of the remaining 29 subjects made a total of eight errors on five experimental words (four subjects made one error and two subjects made two errors; the highest number of subjects making an error on the same item was three of 29). The errors appeared random since they occurred on words that were neither unfamiliar nor infrequent and seemed easily perceivable to other subjects. This was confirmed by the finding that none of the words generated RT outliers for any other subject, except for the word *athlete*, which occurred as an outlier for three subjects. One of the subjects commented that this item sounded like /*aplete*/, which might indicate a perceptual problem. These five experimental items were excluded, in addition to two items which posed perceptual problems (*chlorine*, *factor*) and one item which 20% of the subjects did not seem to know (*despot*); resulting in a total of 42 words for final analysis.

Appendix 2

Range of RT values per stimulus item (29 subjects)

Stimulus	Minimum RT	Maximum RT	Range (max RT − min RT)	Median RT
algae	0·904	1·417	0·513	1·086
angel	0·738	1·158	0·420	0·909
ankle	0·798	1·295	0·497	1·073
barrel	0·901	5·502[a]	4·601	1·166
bottom	0·820	1·295	0·475	1·037
brother	0·707	1·064	0·357	0·892
burden	0·944	1·633	0·689	1·093
comrade	0·976	1·570	0·594	1·171
elbow	0·845	1·303	0·458	1·088
future	0·810	1·730	0·920	1·037
harness	0·881	1·324	0·443	1·086
knowledge	0·884	1·394	0·510	1·045
level	0·765	1·519	0·754	1·015
meadow	0·809	1·262	0·453	1·002
merit	0·942	1·319	0·377	1·134
minute	0·643	1·085	0·442	0·960
music	0·712	1·214	0·502	0·934
ocean	0·715	1·148	0·433	0·950
peasant	0·827	1·997	1·170	1·006
pigeon	0·684	1·189	0·505	0·930
plaza	0·952	1·602	0·650	1·172
pleasure	0·696	1·247	0·551	0·971
poem	0·742	2·472	1·730	0·957
portion	0·915	1·752	0·837	1·145
prairie	0·798	1·439	0·641	1·045
process	1·021	1·339	0·318	1·161
purpose	0·803	1·421	0·618	1·006
refuse	0·974	1·707	0·733	1·239
salad	0·974	1·616	0·642	1·127
savage	0·964	1·720	0·756	1·189
shadow	0·838	1·334	0·496	1·005
shepherd	0·859	1·242	0·383	1·053
signal	1·013	1·358	0·345	1·141
temper	0·862	1·399	0·537	1·024
timber	0·803	1·272	0·469	1·015
value	0·915	1·409	0·494	1·146
venom	0·891	1·474	0·583	1·129
vigil	0·890	2·160	1·270	1·098
village	0·803	1·464	0·661	1·039
visit	0·745	1·302	0·557	0·966
whisper	0·780	1·277	0·497	0·983
woman	0·582	1·249	0·667	0·922

[a] RT outlier (> 3 SD above or below individual mean RT). Median RT: *M* = 1·051; range = 0·892–1·239; SD = 0·087.