Available online at www.sciencedirect.com
SCIENCE DIRECT®
Brain and Language 90 (2004) 160–169
Brain and Language
www.elsevier.com/locate/b&l

The word frequency effect in picture naming: Contrasting two hypotheses using homonym pictures

Keren B. Shatzman[a,*] and Niels O. Schiller[a,b]

[a] *Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands*
[b] *Department of Neurocognition, Faculty of Psychology, University of Maastricht, The Netherlands*

Accepted 2 December 2003
Available online 31 January 2004

Abstract

Models of speech production disagree on whether or not homonyms have a shared word-form representation. To investigate this issue, a picture-naming experiment was carried out using Dutch homonyms of which both meanings could be presented as a picture. Naming latencies for the low-frequency meanings of homonyms were slower than for those of the high-frequency meanings. However, no frequency effect was found for control words, which matched the frequency of the homonyms' meanings. Subsequent control experiments indicated that the difference in naming latencies for the homonyms could be attributed to processes earlier than word-form retrieval. Specifically, it appears that low name agreement slowed down the naming of the low-frequency homonym pictures.

Keywords: Speech production; Lexical access; Phonological encoding; Homonyms; Lemma–lexeme distinction

1. Introduction

The word frequency effect in speech production refers to the finding of Oldfield and Wingfield (1965) that pictures with high-frequency (HF) names (such as *chair*) are named faster than pictures with low-frequency (LF) names (such as *syringe*). A study by Jescheniak and Levelt (1994) provided evidence for the claim that the word frequency effect is due to accessing the phonological forms of words. Though this claim has been generally accepted, the views diverge on the question of which processes are involved in retrieving the phonological form of a word.

In most models of speech production (e.g., Dell, 1986; Jescheniak & Levelt, 1994; Levelt, Roelofs, & Meyer, 1999; Roelofs, 1992) lexical access is assumed to proceed in two steps. First, activation spreads from a conceptual representation (the *lexical concept*) to a semantically appropriate item in the lexicon. This item is referred to as *lemma* and the process as *lemma selection*. It is at the lemma level that syntactic properties of a word (such as whether it is a noun or verb, whether it has masculine or feminine gender, etc.) are activated and can be retrieved. Note that lemmas contain no information regarding a word's phonology. It is only in the second step of lexical access that a word's phonological form, including its segmental content and its metrical properties are retrieved. The phonological representation of the word is referred to as *word form* or *lexeme* and the process of accessing it as *word form* or *lexeme retrieval*.

Recently, it has been argued that it is not necessary to postulate lemma representations that mediate between the semantic-conceptual representations and the phonological form. According to the model proposed by Caramazza (Caramazza, 1997; Caramazza & Miozzo, 1997, 1998), lexical-semantic representations directly activate word-form representations. In this model, called the Independent Network (IN) model, semantic, syntactic, and form representations of a word are independently stored in separate networks. An item's syntactic properties are accessed in the syntactic network in parallel to the lexeme retrieval in the word-form network.

* Corresponding author. Fax: +31-24-3521-213.
E-mail address: keren.shatzman@mpi.nl (K.B. Shatzman).

0093-934X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0093-934X(03)00429-2

The difference in architectures between two-step models and the IN model becomes apparent in the way homophones are represented. Homophones are words that have the same phonology but differ in meaning. In two-step models this property is realized in that homophones share the same word-form representation, but because they have different meanings and different syntactic properties (e.g., "the bear" vs. "to bear"), they have different lemmas. In the IN model, on the other hand, each word, homophonic or non-homophonic, is represented independently. Following Caramazza, Costa, Miozzo, and Bi (2001), we will distinguish between the "shared representation" (SR) assumption and the "independent representation" (IR) assumption.

One consequence of the difference in assumptions regarding how homophones are represented relates to the word frequency effect. If the word frequency effect is at the word-form level, then the SR and IR assumptions carry different predictions with them. Assuming that homophones have a shared representation, the speed of accessing the lexeme is determined by the sum of the frequencies of all meanings of the homophone (cumulative frequency). For example, the speed of accessing the word form /bɛr/ would be determined by the sum of the frequencies of all meanings of the word. Under the SR assumption then, despite the fact that the noun "*bear*" occurs less often than the verb "*bear*," retrieval of this word's lexeme should take as long as retrieval of the verb. In contrast, under the IR assumption, speed of accessing the lexeme is determined by the frequency of each individual meaning.

Experimental research of this issue has produced contradicting evidence. Jescheniak and Levelt (1994) had Dutch–English bilinguals produce Dutch translations of English words. Some items, when translated, resulted in a LF Dutch word that had a HF homophone twin. The study also included two types of control words. The first type were LF non-homophonic words that had the same frequency as the LF meaning of the homophone. The second type were HF non-homophonic words that had the same frequency as the cumulative frequency of the homophone. The results showed that response latencies for producing the homophones (in their LF meaning) were shorter than response latencies for producing the LF control words. Moreover, the latencies for the homophones resembled the response latencies for producing the HF control words that were frequency-matched to the cumulative frequency of the homophone. These results indicate that it is the cumulative frequency that determines the accessing speed to the phonological form of the homophone. This will be referred to as the homophone *cumulative-frequency effect*.

Recently, Jescheniak and Levelt's (1994) results have been questioned. In a picture-naming experiment, Caramazza et al. (2001) found no evidence for a homophone cumulative-frequency effect. Participants had to name pictures that had LF names, which were homophonic to HF words (e.g., *nun* and *none*). The results showed that response latencies to these pictures were similar to response latencies to LF non-homophonic pictures whose names were frequency-matched to the individual (LF meaning of) homophone names. Furthermore, non-homophonic pictures whose names were frequency-matched to the homophone cumulative-frequency were named significantly faster than the homophone pictures. These findings are at odds with the SR assumption, which predicts that the cumulative frequency of both homophone meanings determines the accessing speed of its word form. Furthermore, using a translation task with English–Spanish bilinguals, Caramazza et al. could not find any evidence for the homophone cumulative-frequency effect. However, various properties of the stimuli used by Caramazza et al. and the control tasks in this study differed from those used by Jescheniak and Levelt (for a detailed discussion see Jescheniak, Meyer, & Levelt, 2003). It is possible that the failure of Caramazza et al. to find a homophone cumulative-frequency effect was due to these methodological factors.

The absence of conclusive evidence served as impetus for the current study, which is designed to contrast the predictions of the SR and IR assumptions. Using a picture-naming task, the time it takes to produce the name corresponding to the LF meanings of homophones will be compared to the time it takes to produce their HF twins. Consider, for instance, the English word *bat*, which could refer to a *baseball bat* or a *flying mammal*. Both noun meanings could be presented as a picture. Suppose further that one meaning is much more frequent than the other. Other things being equal, the IR assumption predicts that the picture depicting the more frequent meaning should be named quicker than the picture that depicts the less frequent meaning because different word forms have to be retrieved. The SR assumption, on the other hand, predicts that producing "bat" in either meaning would take the same amount of time because the same word form has to be retrieved. Thus, using such differentially frequent pictorial homophones the predictions of the SR and IR assumptions can be tested.

A list of pictorial, semantically non-related Dutch homophones was constructed. All homophones were also homonyms, i.e., had identical orthography. Because objective frequency counts, such as the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995) do not provide separate counts for homonyms, subjective frequency ratings (Experiments 1a and 1b) were carried out to determine the frequency of each meaning.

2. Experiments 1a and 1b: Subjective frequency rating

Previous studies have shown that subjective frequency ratings correlate highly with objective frequency counts (e.g., Carroll, 1971; Shapiro, 1969). This fact was utilized to obtain estimates of separate meaning frequency for the homonyms, in the following way: participants rated words, which were either homonyms of the same grammatical class (words like "bat"), homonyms of different grammatical classes (words like "bear", which could be a noun or a verb) or non-homonyms. For homonyms of different classes and non-homonyms CELEX provides frequency counts. Therefore, for these items, a regression line could be calculated which describes the relationship between the objective and subjective frequency. Using the same regression line, the predicted objective frequency could be calculated for each meaning of a (same class) homonym. To that end the subjective frequency of each meaning is entered into the equation that describes the regression line.

To increase the reliability of the ratings, we performed two separate rating experiments, using different techniques. In Experiment 1a, which was based on a method used by Griffin (1999), ratings were given on an *open-end* scale. Presenting pictures depicting the meaning of the word disambiguated the homonyms. In Experiment 1b, based on de Jong (2002), a *closed-end* scale was used. A disambiguating word was given for the homonym words to clarify which meaning should be rated. The results of these two experiments were then integrated to give one frequency estimate for each meaning of the homonym.

2.1. Method

2.1.1. Participants

Forty-one native speakers of Dutch recruited from the pool of participants of the Max Planck Institute for Psycholinguistics took part in the experiment (25 in Experiment 1a and sixteen in Experiment 1b). All participants were students at the University of Nijmegen. They received Dfl. 8.50 for their participation.

2.1.2. Materials

Thirty-nine Dutch homonyms were selected. Of these, 12 homonyms were of different grammatical class. The other 27 were same-class homonyms. Both readings of each homonym could be presented as a picture. The two readings were semantically unrelated to each other (see Appendix A for a complete list of the homonyms). In addition to the homonyms, we tested 150 non-homophonic words. These included 50 LF words (less than 10 occurrences per million according to CELEX), 50 HF words (more than 50 occurrences per million), and 50 medium-frequency words (between 10 and 50 occurrences per million).

In Experiment 1a, for each word and each reading of the homonyms, a picture depicting its meaning was either chosen from the picture database at the Max Planck Institute or created using the Adobe Illustrator drawing program. Altogether there were 228 pictures in the experiment. Picture size was scaled to fit in a 52×52 mm frame. The pictures were then printed on A4 paper so that each page contained six pictures. Underneath each picture the picture's name appeared in lower case Times New Roman 24-point typeface. Next to the picture's name a line was drawn on which participants were to write that word's estimated frequency.

In Experiment 1b, each word appeared next to a seven-point scale. A short disambiguating description accompanied words with more than one meaning.

2.1.3. Design

Two lists of items were constructed in such a way that different meanings of each homonym appeared on different lists. Each list consisted of 114 items. In Experiment 1a, each participant saw both lists, and the order of list presentation was counterbalanced across participants. In Experiment 1b, each participant saw only one list. Within each list, word order was a different random sequence for each participant.

2.1.4. Procedure

Participants were asked to rate how often they typically said a word, using a particular meaning. In Experiment 1a, they were told to rate meaning use relative to the other items in a list:

> "Imagine that you have given the word *milk* a value of 100. Subsequently, we would like you to estimate how often you use other words in speaking. If a word seems twenty times as frequent as another, you would give it a number twenty times as large (thus, in this example, 2000). If, on the other hand, the word seems only half as frequent, give it a number half as large (in this example 50)... You may use whole numbers, fractions or decimals, but not negative numbers. Only use a zero if you have never used a word."

The word in the instructions (*milk*) was only used as an example, and participants were allowed to set their own anchor. Every participant was presented with six practice items, selected to represent a wide frequency range. Pictures appeared next to all the words, so that the intended meaning would be clear. The participants wrote down on paper next to each item their estimation of that item's frequency. The results of each participant's ratings were normalized, so that for each word that a participant rated a z-score based on the deviation from the participant's mean rating was calculated. The median z-score for each item was calculated by taking the median score across participants.

In Experiment 1b, participants were asked to indicate on a seven-point scale how often they thought a word

was used. They were instructed to take into consideration the disambiguating description, whenever this was available, and rate only the relevant meaning in that case. The mean score was calculated per item.

2.2. Results and discussion

A regression analysis was performed on the non-homophonic words, testing the linear relationship between the median *z*-score (Experiment 1a) or mean score (Experiment 1b) of each word and its log CELEX frequency. That is, testing the models:

$$\text{median} = a^*\log(\text{CELEX frequency}) + b^1 \quad \text{(Experiment 1a)}$$
$$\text{mean} = a^*\log(\text{CELEX frequency}) + b^1 \quad \text{(Experiment 1b)}$$

In the regression analysis the models proved to be significant (Experiment 1a: $F(1,149) = 47.3$, $p < .001$; Experiment 1b: $F(1,149) = 133.2$, $p < .001$). In Experiment 1a, a significant correlation was found between the median *z*-score of each word and its log CELEX value ($r(150) = .49$, $p < .001$). The correlation between mean score and log CELEX values in Experiment 1b was also significant ($r(150) = .69$, $p < .001$).

With the coefficients of the regression line and the subjective ratings of each word we could calculate the predicted log frequencies for each word, using the formulae:

$$\text{predicted log(CELEX frequency)} = (\text{median} - b)/a \quad \text{(Experiment 1a)}$$
$$\text{predicted log(CELEX frequency)} = (\text{mean} - b)/a \quad \text{(Experiment 1b)}$$

The results of both experiments were then combined by averaging the predicted log CELEX from both experiments. For the non-homonyms, the average predicted log CELEX values were nearly identical to the observed CELEX values (means 3.08 and 3.06, respectively). In a paired two-samples *t* test the two values did not differ significantly from each other ($t(149) < 1$). The correlation between the average predicted log CELEX and the observed log CELEX was highly significant ($r(150) = .62$, $p < .001$).

Next, we looked at the different class homonyms. Because they belong to different grammatical categories (for instance, "bear" as a noun or a verb), these items have separate CELEX frequency for each reading of the homonym. The separate CELEX frequencies allowed us to compare predicted CELEX values with real CELEX values for these homonyms. The subjective rating scores (i.e., median *z*-scores in Experiment 1a and mean scores in Experiment 1b) were entered into the regression model (using the coefficients from the regression analysis of the non-homonym words) and predicted log CELEX values were calculated and averaged. As with the non-homonyms, the average predicted log CELEX values were nearly identical to the observed CELEX values (means 2.67 and 2.60, respectively) and the two did not differ significantly from each other ($t(23) < 1$). The correlation between the average predicted log CELEX and the observed log CELEX was highly significant ($r(24) = .70$, $p < .001$).

For the same-class homonyms CELEX only provides a frequency value aggregated over meanings. That is, the CELEX value for "bat" is the sum of the frequency of baseball bat and flying mammal. The subjective rating scores of the same-class homonyms were entered into the regression model (as was done with the different class homonyms) and the predicted CELEX value for each meaning of each word was calculated. Then, the predicted values of both meanings of each homonym were added up to yield the sum of the average predicted CELEX values. The sum of the average predicted log CELEX per homonym word was compared to the observed log CELEX. The predicted and observed log CELEX frequencies were not significantly different (mean sum of predicted log CELEX: 3.23, mean observed log CELEX: 3.11, $t(26) = 1.17$, *n.s.*).

Thus, the average predicted log CELEX proved to be an accurate measurement in predicting frequency both for the non-homonyms and the homonyms. Therefore, for the purpose of selecting items for the picture-naming experiment, the average of predicted CELEX frequency seemed to be a reliable and accurate frequency measurement. We will refer this measurement as *rated frequency*.

3. Experiment 2: Picture naming

The current study is designed to contrast the predictions of the SR and IR assumptions. According to the SR hypothesis, homophones share their lexeme, that is, their phonological word-form representation. In contrast, the IR hypothesis claims that each homophone has a separate phonological representation. Assuming that the speed of accessing the phonological form is determined by the threshold activation of the phonological form, the SR hypothesis predicts that homophones will have the same accessing speed for both meanings. In other words, because the lexeme is the same, the speed of accessing it will be the same.

The IR hypothesis, on the other hand, predicts that the lexeme of the more frequent meaning of the homophone will be accessed faster than the lexeme of the less frequent meaning. According to this account, homophones are just like non-homophonic words. The lexemes of HF words are accessed faster than those of LF words, regardless of whether or not the word is a homophone.

[1] Because word frequency counts are logarithmic, the log-transform of frequency is used to describe a linear relationship with the subjective frequency. The median and the mean are a function of CELEX log frequency, with *a* and *b* as the coefficients of that linear function.

3.1. Method

3.1.1. Participants

Thirty paid participants recruited from the Max Planck Institute's pool of participants took part in the experiment.

3.1.2. Materials

There were four experimental item sets: Hom-HF, Hom-LF, control-HF, and control-LF. Hom-HF items were pictures that depicted the dominant meaning of each homonym. Hom-LF pictures depicted the subordinate meaning. For each of the homonym pairs the rated frequency of the Hom-HF name exceeded the rated frequency of the Hom-LF name by both more than 13 occurrences per million and by at least a factor of two. Control-HF items were pictures with non-homophonic names whose rated frequency matched that of the Hom-HF items. Control-LF items had non-homophonic names whose rated frequency matched that of the Hom-LF items. The four sets included only items with morphologically simple names. Control items were selected such that there would be no systematic difference in word-initial manner of articulation between the homonym and control items. Furthermore, control-LF and control-HF items were matched for number of syllables and number of phonemes. With the exception of two items, these two sets had a perfect match of word onset (see Appendix B for a complete list of the items).

Each of the experimental item sets included 16 pictures. Mean rated frequencies (per million) for the different sets were 8 (Hom-LF), 10 (control-LF), 122 (Hom-HF), and 101 (control-HF). In addition to the experimental items, 16 filler items were selected. The names of the filler items were all non-homophonic and shared word-length and frequency characteristics of the experimental items.

Finally, there were 10 practice pictures. Altogether there were 90 pictures in the experiment. All pictures were line drawings of objects, selected from the picture database at the Max Planck Institute or created using the Adobe Illustrator drawing program.

3.1.3. Design

Participants were randomly assigned to one of two conditions, so that each participant was only exposed to one meaning of the homonym. Each participant received 8 Hom-LF items, 8 Hom-HF items, 16 control-LF items, and 16 control-HF items, as well as 16 filler items. Each item was presented three times, giving a total of 192 trials (preceded by 10 practice trials). The trials were divided to three blocks and in each block a given picture appeared once. For each of the two conditions, four pseudo-randomized trial sequences were constructed, with the constraints that (a) homonyms did not appear on consecutive trials; (b) no item would be preceded by a phonologically or semantically related item; and (c) repeated presentations of any experimental item were separated by at least 20 intervening trials.

3.1.4. Apparatus

The experiment was run on a Hermac 486 computer. The pictures were presented on a NEC Multisync II screen. Participants responded into a Sennheiser ME400 microphone. The trial sequencing was controlled by NESU (Nijmegen Experimental Set-Up) and naming latencies were measured using a voice key. All sessions were taped with a Sony DTC55 DAT recorder.

3.1.5. Procedure

Participants were tested individually in a sound-attenuated booth. The pictures were presented as white line drawings on a black background. Display size of the pictures was scaled to fit into a 74×74 mm frame. Viewing distance was approximately 60 cm.

The experiment began with a learning phase. Participants were exposed to the pictures, one at a time, and asked to name them with the most appropriate name they could think of. After they gave their response, the name that was to be used in the experiment appeared on the screen, under the picture, in lower case Arial 36-point typeface. If that name deviated from their original response, they were instructed to read the name aloud. The picture and the name stayed on the screen for at least 2 seconds and participants were asked to study these carefully in order to know which name to use for any given picture. The experimenter noted all alternative names.

After the learning phase the picture-naming experiment started. At the beginning of each trial, a fixation point was presented in the center of the screen for 500 ms. Following a pause of 500 ms, the target picture appeared on the screen and remained visible until the voice key was activated. However, if no response was registered within 2000 ms, the picture disappeared anyway and after 1500 ms the next trial began. The experiment started with a short training phase of 10 practice items. The participants' responses were monitored by the experimenter and scored for correctness.

Participants were instructed to name the pictures as quickly as possible, without making errors. They regularly received feedback on their speed: every 20 trials their average reaction time appeared on the screen and they were asked to write it down on a piece of paper. This had the purpose of speeding participants up. The feedback pause also allowed participants to rest.

3.1.6. Analysis

Responses were scored as errors and were excluded from the analysis in case (a) the target picture name was

not produced; (b) the voice key was triggered by a nonverbal sound; (c) a verbal disfluency occurred or an utterance was repaired; or (d) the speech onset latency exceeded 2000 ms. Responses were also excluded if their latencies deviated by more than two standard deviations from a participant's or an item's mean latency.

Averaged reaction times and errors were submitted to two separate analyses of variance (ANOVAs), with subjects (F_1) and items (F_2) as random variables. Statistical analyses involved two fixed variables: frequency (low vs. high) and presentation (first vs. second vs. third). Because they involved different designs, control and homonym items were analyzed separately. In the by-subject analysis of the control items, frequency and presentation were treated as within-subject variables. In the by-item analysis, frequency was treated as a between-item variable and presentation as a within-item variable. For the by-subject analysis of the homonym items, frequency was treated as a between-subject variable and presentation as a within-subject variable. In the by-item analysis, both frequency and presentation were treated as within-item variables.

3.2. Results and discussion

Overall, on the basis of the above-mentioned criteria, 314 observations (7.3%) were marked as errors. Outliers accounted for an additional 6.2% of the data, equally distributed over the four item groups. Mean response latencies for the control items sets are displayed in Fig. 1.

Overall, HF items were named slightly faster than LF items. This effect was significant in the by-subject analysis ($F_1(1,29) = 13.76$, $MS_e = 391$, $p < .01$), but not in the by-item analysis ($F_2(1,30) < 1$). With repeated presentation, responses became faster, averaging 631, 587, and 578 ms, on the first, second, and third presentation, respectively. This yielded a significant presentation effect ($F_1(2,58) = 73.42$, $MS_e = 716$, $p < .001$; $F_2(2,60) = 100.12$, $MS_e = 297$, $p < .001$). As Fig. 1 shows, the frequency effect decreased with repetition and disappeared completely in the third presentation. This interaction between frequency and presentation was significant in the by-subject analysis ($F_1(2,58) = 3.64$, $MS_e = 331$, $p < .05$), but not in the by-item analysis ($F_2(2,60) = 1.68$, $MS_e = 498$, *n.s.*). In the analysis of error rates, no significant effects were obtained. Average naming latencies for the homonyms are presented in Fig. 2.

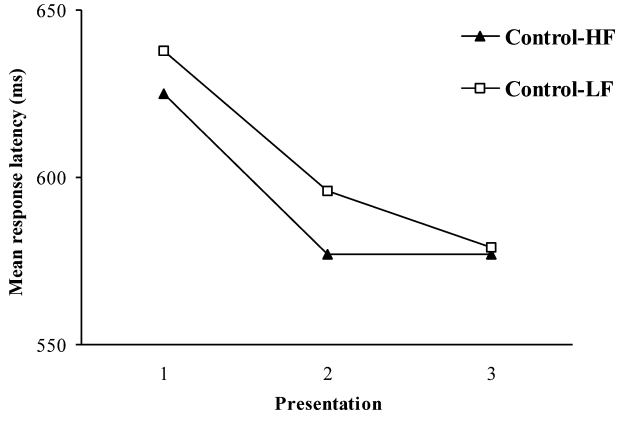

Fig. 1. Mean speech onset latencies for the control items in Experiment 2.

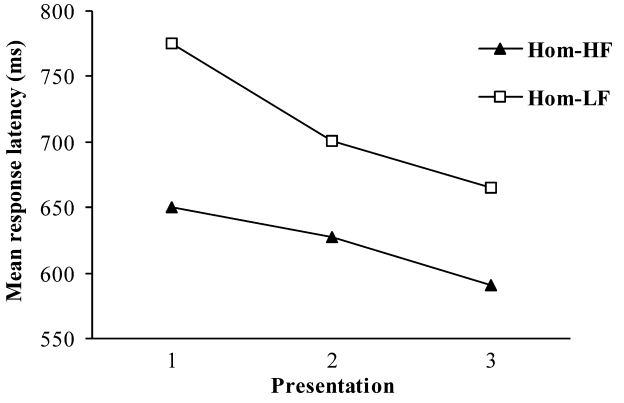

Fig. 2. Mean speech onset latencies for the homonym items in Experiment 2.

Overall, naming latencies for pictures with Hom-LF names were 87 ms slower than naming latencies for pictures with Hom-HF names. This effect was significant both by subjects and by items ($F_1(1,58) = 44.06$, $MS_e = 8308$, $p < .001$; $F_1(1,15) = 14.16$, $MS_e = 14,614$, $p < .01$). There was a significant repetition effect, with responses becoming faster with repeated presentation—presentations 1 through 3: 706, 663, and 627 ms, respectively ($F_1(2,116) = 67.94$, $MS_e = 1664$, $p < .001$; $F_2(2,30) = 13.36$, $MS_e = 1006$, $p < .001$). The magnitude of the frequency effect decreased from 125 ms on the first presentation to 74 on the second and third presentation, yielding an overall significant interaction between frequency and presentation ($F_1(2,116) = 6.31$, $MS_e = 1664$, $p < .01$; $F_2(2,30) = 53.39$, $MS_e = 1153$, $p < .001$).

Error rates for the homonym item sets are displayed in Fig. 3. Participants made more errors on Hom-LF trials than on Hom-HF trials—13.6 and 6.5%, respectively ($F_1(1,58) = 10.19$, $MS_e = 220$, $p < .01$; $F_2(1,15) = 8.47$, $MS_e = 142$, $p < .05$). There was a significant effect of presentation, with error rates dropping from 14% on the first presentation to 9% on the second and 8% on the third presentation ($F_1(2,116) = 6.84$, $MS_e = 94$, $p < .01$; $F_2(2,30) = 6.98$, $MS_e = 48$, $p < .01$).

The results of Experiment 2 showed a small and weak frequency effect for the control items, with control-HF items produced slightly faster than control-LF items. The effect was not reliable, being significant by

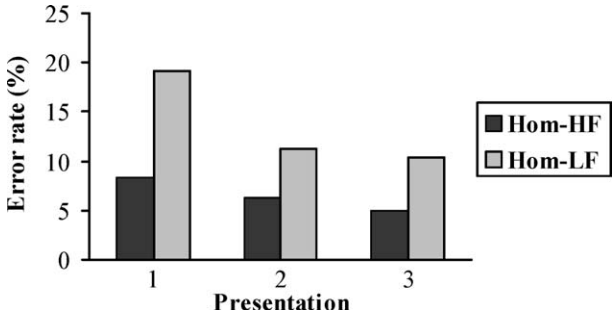

Fig. 3. Error rates for the homonym items. The corresponding values for the first, second, and third presentations, respectively, were 19, 11, and 10% for Hom-LF, and 8, 6, and 5% for Hom-HF.

subjects but not by items. In contrast, the homonym items produced a large and reliable frequency effect, with Hom-LF items named much slower than Hom-HF items. Prima facie, the difference in naming latencies between the high- and low-frequency homonyms constitutes supporting evidence for the IR hypothesis, namely, that homonyms have separate phonological representations. However, the magnitude of the homonym frequency effect, compared to the effect found for the frequency-matched control items, indicates that Hom-LF and Hom-HF items probably differed in other aspects besides frequency. The presence of a frequency effect in the error data supports this argument. One possibility, for instance, is that the observed frequency effect was caused by differences between the stimuli picture sets. This possibility was examined in Experiment 3a.

4. Experiment 3a: Picture recognition

In order to name a picture, participants need to first recognize the visual object and access the appropriate lexical concept. Experiment 3a was designed to examine whether or not these early perceptual and conceptual processes contributed to the results obtained in Experiment 2. Immediately after the picture-naming experiment, the same participants performed the following verification task: A picture was presented very briefly on the screen, followed by a congruent or non-congruent word. The participants' task was to decide whether or not the word denoted the object in the picture and to push a *yes* or *no* button accordingly. Pictures in which the depicted object is easily recognized and prototypical of that lexical concept should elicit shorter reaction times compared to pictures that are difficult to recognize or atypical for that concept. Thus, if the frequency effect observed in Experiment 2 arose from differences in object recognition or conceptual access latencies, this should also be reflected by reaction times in the verification task.

4.1. Method

4.1.1. Participants

The participants were the same as in Experiment 2.

4.1.2. Materials

The pictures used in Experiment 2 were intermixed with a new set of 32 filler items. The control and homonym items of Experiment 2 were always *yes* trials, while the fillers from Experiment 2 and the new set of filler items were always *no* trials. Therefore, there was an equal number of *yes* and *no* responses. Finally, the practice items of Experiment 2 served again as practice trials, one half of them presented in *yes* trials and one half in *no* trials.

4.1.3. Design

As in Experiment 2, there were two conditions, so that each participant was only exposed to one meaning of the homonym, that is, the meaning he or she received in Experiment 2. Each item was presented only once. For each of the two conditions, two pseudo-randomized trial sequences were constructed, with the constraints that homonyms did not appear on consecutive trials and no experimental item would be preceded by a phonologically or semantically related item.

4.1.4. Procedure

Each participant was tested individually. All visual stimuli were presented centered on the screen. The words were displayed in lower case Arial 36-point typeface. Two push buttons were used, one for the *yes* response and one for the *no* response. The *yes* response was always assigned to the participant's dominant hand.

At the beginning of each trial, a fixation point was presented in the center of the screen for 500 ms. Following a pause of 500 ms, a picture appeared on the screen for 150 ms. Immediately after that, a word was displayed for 1500 ms. The word display stopped when the participant pressed a button. However, if no response was registered within 1500 ms, the picture disappeared and after 1500 ms the next trial began. The experiment started with a short training phase of 10 practice items. After a short pause, the 96 test items were presented.

Participants were instructed to respond as quickly as possible, without making errors. They regularly received feedback on their speed: every 20 trials their average reaction time appeared on the screen and they were asked to write it down.

4.1.5. Analysis

All incorrect push-button responses and latencies exceeding 1500 ms were treated as errors and excluded from the data. Responses were also excluded if their latencies deviated by more than two standard deviations

from a participant's or an item's mean latency. The main data analyses were carried out on the experimental items, that is, the items requiring a *yes* response. As in Experiment 2, control and homonym items were analyzed separately.

4.2. Results and discussion

On the basis of the above-mentioned criteria, 45 observations (3.1%) were marked as errors. Outliers accounted for an additional 7.2% of the data, equally distributed over the four item groups. Averaged reaction times and errors were submitted to two separate analyses of variance (ANOVAs), with subjects (F_1) and items (F_2) as random variables.

Mean reaction times, standard deviations, and error rates for each item set are shown in Table 1.

While reaction times for control-LF and control-HF pictures did not differ statistically (both Fs < 1), verification was performed more rapidly to Hom-HF items than to Hom-LF items ($F_1(1,58) = 4.66$, $MS_e = 4538$, $p < .05$; $F_2(1,15) = 14.85$, $MS_e = 970$, $p < .01$). The error rate analysis did not yield a significant frequency effect (control items: $F_1(1,29) = 1.40$, $MS_e = 17$, *n.s.*; $F_2(1,30) = 2.0$, $MS_e = 6$, *n.s.*; homonyms: both Fs < 1).

The results of Experiment 3a showed no difference in verification latencies for the control items. This suggests that LF and HF control items in Experiment 2 did not differ in early perceptual and conceptual processing. For the homonym pictures, the verification task revealed a 40-ms frequency effect. Given that pictures were presented for 150 ms, a time-frame that only allows superficial visual processing and accessing the lexical concept (Levelt, Praamstra, Meyer, Helenius, & Salmelin, 1998; Thorpe, Fize, & Marlot, 1996), participants could not have retrieved a picture's name before it appeared on the screen. Therefore, the frequency effect in the verification task could not be due to phonological retrieval.

The most obvious explanation for the frequency effect in the verification task is that the dominant meaning of the homonym—the Hom-HF meaning—is activated faster and stronger by the presented word than the subordinate meaning. Consequently, verification in Hom-HF trials can occur as soon as the meaning of the presented word has been accessed. In Hom-LF trials, in contrast, verification can only occur after the subordinate meaning of the presented word has been accessed, resulting in slower verification times. Thus, the frequency effect in the verification task indicates that homonym names corresponded better to Hom-HF pictures than to Hom-LF pictures. This argument leads to the prediction that name agreement would be higher for Hom-HF pictures than for Hom-LF pictures. This prediction was examined in Experiment 3b.

Table 1
Mean reaction time (in milliseconds), standard deviations, and error rates (in percentages) in Experiment 3a

Condition	Mean RT	*SD*	Error rate
Control-HF	453	92	1.9
Control-LF	450	84	3.1
Hom-HF	449	84	3.8
Hom-LF	489	98	5.0

5. Experiment 3b: Picture–name agreement

Before the picture-naming experiment, participants went through a learning phase, in which they were exposed to the pictures, one at a time, and were asked to name them with the most appropriate name they could think of. Analyzing the spontaneous naming responses, which were elicited in this learning phase, could, therefore, reveal differences in name agreement between the picture sets.

5.1. Method

5.1.1. Participants

The participants were the same as in Experiment 2.

5.1.2. Materials

The picture stimuli were the same as in Experiment 2.

5.1.3. Design

As in Experiment 2, there were two conditions such that each participant was only exposed to the picture depicting one meaning of the homonym. For each of the two conditions, a trial sequence was constructed such that homonyms did not appear on consecutive trials.

5.1.4. Procedure

Participants were exposed to the pictures, one at a time, and asked to name them with the most appropriate name they could think of.

5.1.5. Analysis

The experimenter noted all responses deviating from an item's designated name. The percentages of deviating responses were submitted to ANOVAs, with subjects (F_1) and items (F_2) as random variables. As in Experiment 2, control and homonym items were analyzed separately.

5.2. Results and discussion

Mean percentages of deviating responses and standard deviations are shown in Table 2.

The difference between the mean percentage of deviating responses for control-LF items and control-HF items was not significant ($F_1(1,29) = 3.78$, $MS_e = 38$,

Table 2
Mean percentages (and standard deviations) of deviating responses in spontaneous naming

Condition	Mean percentage of deviating responses	*SD*
Control-HF	9.3	16
Control-LF	13.2	18
Hom-HF	27.9	36
Hom-LF	76.4	29

$p = .06$; $F_2(1, 30) < 1$). Hom-LF items were named with a deviant name almost 50% more often than Hom-HF items, yielding a significant effect of frequency ($F_1(1, 58) = 123.84$, $MS_e = 278$, $p < .001$ and $F_2(1, 15) = 15.27$, $MS_e = 1232$, $p < .01$).

The results of Experiment 3b show that, in spontaneous naming, LF control items were named with an alternative name approximately as often as HF control items. In contrast, there was a huge difference between Hom-LF and Hom-HF items, with Hom-LF items, more often than not, being named with an alternative name. This result suggests that naming latencies were slower for Hom-LF items not because of their lower frequency but because they had more alternative names. In fact, when the percentage of deviating names (in spontaneous naming) and the rated frequency of the homonym words were entered into a regression model as predictors for naming latencies, percentage of deviating names explained a larger share of the variance ($R^2 = .61$, $p < .001$) than rated frequency ($R^2 = .42$, $p < .001$). Furthermore, in a step-wise regression, the inclusion of frequency did not add any significant explanatory power to the model (Model 1, percentage of deviating names: $R^2 = .61$; Model 2, percentage of deviating names *and* rated frequency: $R^2 = .65$). Moreover, for a subset of the homonyms, for which the percentage of deviating names for LF and HF items was similar, there was a small difference in mean naming latencies (672 ms and 683 ms for Hom-HF and Hom-LF items, respectively), but this difference did not yield a reliable effect of frequency ($F_1(1, 58) = 1.74$, $MS_e = 8353$, *n.s.*; $F_2(1, 6) < 1$).

The results of these post hoc analyses, together with the results from Experiments 3a and 3b, indicate that the difference in naming latencies observed for the homonyms in picture naming (Experiment 2) probably was not driven by the difference in frequency but by the difference in name agreement. This suggests that the effect did not arise in the process of phonological retrieval. Rather, the effect could have taken place during earlier processes. The fact that many Hom-LF items had near-synonyms could by itself slow down the naming of those Hom-LF pictures, compared to Hom-HF pictures. Several studies showed that objects with low name agreement take longer to name than objects with high name agreement (Lachman, 1973; Lachman, Shaffer, & Hennrikus, 1974; Vitkovitch & Tyrell, 1995). In the case of the Hom-LF items, the situation is more extreme because the Hom-LF names had more dominant counterparts (e.g., *slot* in the meaning of "castle" has a higher frequency counterpart *kasteel* with a very similar meaning).

6. General discussion

The experiments reported here investigated the lexical representation of homonyms. In Experiments 1a and 1b, subjective ratings were used to determine the frequency of each meaning of a set of homonyms. The homonyms for which one meaning was much more frequent than the other were selected as items for the picture-naming experiment (Experiment 2). Naming latencies for pictures depicting the low-frequency meaning of the homonym (Hom-LF) were much slower compared to pictures depicting the high-frequency meaning (Hom-HF). This large frequency effect was not found in the naming latencies of pictures whose (non-homonymic) names were frequency-matched to the homonyms. In Experiment 3a, it was shown that verification of Hom-LF pictures was slower than verification of Hom-HF pictures. Experiment 3b, involving spontaneous naming, showed that name agreement was lower for Hom-LF pictures than for Hom-HF pictures. Furthermore, in a regression analysis, it was found that naming latencies were better predicted by the name agreement than by the frequency variable. When frequency was included as a factor in the regression model, the gain in explained variance was not significant. Moreover, when Hom-LF and Hom-HF items were matched on name agreement, the frequency effect disappeared. These results, therefore, lead to the conclusion that the difference in naming latencies between Hom-LF and Hom-HF items is not truly the word frequency effect that is due to accessing the phonological form of a word, but rather reflects differences in other processes, most probably lexical selection. Consequently, these results do not allow one to conclude whether or not homonyms have a shared phonological representation.

The most striking feature of the homonyms was that name agreement was much lower for Hom-LF items than their Hom-HF twins. In fact, this feature might be inherent for homonyms for which there is a large discrepancy between the frequencies of the dominant and subordinate meaning. Name agreement as measured in spontaneous naming, requires participants to name the object in the picture with the most appropriate name they can think of. It is likely that participants do not spontaneously name Hom-LF pictures with the designated homonym name precisely because that word is more commonly used to refer to a different object. In other words, the designated homonym name seems inappropriate to use in referring to the Hom-LF object because that name primarily refers to another object.

This emphasizes a general tendency of natural languages, i.e., to avoid more than one name to refer to a specific object.

Finally, this study provides an elegant and simple method for estimating the frequency of same-class homonyms. As homonyms are widely used in psycholinguistic research, the value of an accurate estimate for meaning frequency can not be overrated.

Appendix A. List of homonyms used in Experiments 1a and 1b

The approximate English translation for both meanings is given in square brackets.

A.1. Different class homonyms

arm [arm, poor], as [ash, axis], bal [ball], been [leg, bone], bos [forest, bunch], bus [bus, bin], das [tie, badger], golf [golf, wave], pad [path, toad], schop [kick, shovel], veer [feather, spring], wortel [carrot, root].

A.2. Same-class homonyms

bank [sofa, bank], blad [leaf, sheet], blik [can, dustpan], bloem [flower, flour], boog [bow, arch], bord [plate, sign], bril [glasses, toilet seat], ezel [donkey, easel], hoorn [horn], kaart [card, map], knoop [button, knot], kop [head, cup], kraan [crane, faucet], kruk [stool, crutch], kwast [brush, tassel], motor [engine, motorcycle], muis [mouse], noot [nut, note], pak [suit, parcel], palm [palm], peer [pear, bulb], riem [belt, oar], schaal [bowl, scale], schrift [writing, notebook], slot [castle, lock], trommel [drum, box], vleugel [wing, grand piano].

Appendix B. List of picture names used in Experiment 2

B.1. Homonyms

bal, been, blik, bloem, bord, bos, bril, bus, hoorn, kaart, knoop, kop, kraan, kwast, pad, slot.

B.2. LF controls

bad [bath], ballon [balloon], beer]bear], berg [mountain], bijl [axe], dolk [dagger], duif [pigeon], hert [deer], kers [cherry], ketel [kettle], koe [cow], koets [carriage], koffer [suitcase], pauw [peacock], slak [snail], tomaat [tomato].

B.3. HF controls

baby [baby], boom [tree], boot [boat], borstel [brush], bureau [desk], dak [roof], duim [thumb], hond [dog], kaas [cheese], kast [closet], kip [chicken], klok [clock], paard [horse], schaar [scissors], taart [cake], trein [train].

References

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

Caramazza, A. (1997). How many levels of processing are there in lexical access. *Cognitive Neuropsychology, 14*, 177–208.

Caramazza, A., & Miozzo, M. (1997). The relation between syntactic and phonological knowledge in lexical access: evidence from the 'tip-of-the-tongue' phenomenon. *Cognition, 64*, 309–343.

Caramazza, A., & Miozzo, M. (1998). More is not always better: A response to Roelofs, Meyer, & Levelt. *Cognition, 69*, 231–241.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1430–1450.

Carroll, J. B. (1971). Measurement properties of subjective magnitude estimates of word frequency. *Journal of Verbal Learning and Verbal Behavior, 10*, 722–729.

de Jong, N. H. (2002). Homonyms in Context. In *Morphological families in the mental lexicon. MPI Series in Psycholinguistics* (Vol. 20). Doctoral Dissertation, University of Nijmegen, pp. 105–146.

Dell, G. S. (1986). A spreading-activation model of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Griffin, Z. M. (1999). Frequency of meaning use for ambiguous and unambiguous words. *Behavior Research Methods, Instruments, and Computers, 31*, 520–530.

Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824–843.

Jescheniak, J. D., Meyer, A. S., & Levelt, W. J. M. (2003). Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 432–438.

Lachman, R. (1973). Uncertainty effects on time to access the internal lexicon. *Journal of Experimental Psychology, 99*, 199–208.

Lachman, R., Shaffer, J. P., & Hennrikus, D. (1974). Language and cognition: Effects of stimulus codability, name-word frequency, and age of acquisition on lexical reaction time. *Journal of Verbal Learning and Verbal Behavior, 13*, 613–625.

Levelt, W. J. M., Praamstra, P., Meyer, A. S., Helenius, P., & Salmelin, R. (1998). A MEG study of picture naming. *Journal of Cognitive Neuroscience, 10*, 553–567.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Brain and Behavioral Sciences, 22*, 1–75.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273–281.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*, 107–142.

Shapiro, B. J. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behavior, 8*, 248–251.

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature, 381*, 520–522.

Vitkovitch, M., & Tyrell, L. (1995). Sources of disagreement in object naming. *Quarterly Journal of Experimental Psychology, 48*, 822–848.