Frequency Effects in the Processing of Chinese Inflection

James Myers*
National Chung Cheng University

Yu-chi Huang
National Taiwan Normal University

Wenling Wang
National Chung Cheng University

*Corresponding author.
Address:
Graduate Institute of Linguistics
National Chung Cheng University
168 University Road, Min-Hsiung
Chia-Yi 62102 Taiwan
Lngmyers@ccu.edu.tw

NOTE: This is the draft of a paper accepted by *Journal of Memory and Language* (as of Dec. 3, 2005). The published version may differ.

Abstract

Chinese inflection differs from that of European languages in that it is fully parsable in the orthography, which raises the possibility that Chinese inflected forms may not show the surface frequency effects found in other languages. Five lexical decision experiments were conducted to examine this issue. They showed that surface frequency did indeed affect reaction times, independent of base frequency and acceptability of the inflected forms, but only if the design of the materials encouraged readers to process base-affix combinations. This suggests that surface frequency effects emerge at a late stage when components are combined into complex words, not during initial contact with the lexicon.

Keywords: surface frequency, base frequency, Chinese, inflection, acceptability, decomposition

Frequency Effects in the Processing of Chinese Inflection

In languages like English and Dutch, it has been shown that lexical decisions for regularly inflected forms matched on base frequency (i.e., the cumulative frequency of all surface forms containing the base morpheme) may be facilitated by higher surface frequency of the inflected forms (Alegre & Gordon, 1999; Baayen, Dijkstra, & Schreuder, 1997; Baayen, Schreuder, De Jong, & Krott, 2002; Bertram, Schreuder, & Baayen, 2000; Burani, Salmaso, & Caramazza, 1984; Katz, Rexer, & Lukatela, 1991; New, Brysbaert, Segui, Ferrand, & Rastle, 2004; Sereno & Jongman, 1997; Taft, 1979). For example, Sereno and Jongman (1997) found that when regularly inflected English words were matched on base frequency, lexical decision times were faster for higher-frequency inflected forms (e.g., *windows* faster than *rivers*). The goal of the present paper, in a nutshell, is to apply the methodology of Sereno and Jongman (1997) to Chinese to determine whether regularly inflected forms in this language also show surface frequency effects.

Aside from adding another language to the list of those tested, there are two major reasons for addressing this issue in Chinese. First, the question motivating the search for surface frequency effects is, of course, whether morphologically complex words can be retrieved from memory as wholes during lexical access, but Chinese turns this question on its head, due to its unusual orthographic system. A Chinese character virtually always represents a single morpheme (and syllable), and word boundaries are not marked, so Chinese readers are forced to work actively to find words through character combination. We know that word composition eventually takes place; belying its reputation as an "isolating" language, in Chinese most words are morphologically complex (Zhou & Marslen-Wilson, 1994), and linguists have developed reliable tests for distinguishing words from phrases (Packard, 2000; Xue, 2001). There is also considerable behavioral evidence for the word level in Chinese reading, including word superiority effects in character and character-component detection tasks (Hung, Tzeng, & Ho, 1999) and word frequency effects in lexical decision tasks (see reviews in Myers, in press; Taft, Liu, & Zhu, 1999; Zhou & Marslen-Wilson, 2000). Thus in Chinese reading the central issue is not decomposition but composition: How do readers derive words from character strings?

The second reason why it is particularly interesting to look at Chinese inflection is that it is often assumed not to exist; this view is reflected, for example, in the title of Li, Bates, and MacWhinney's (1993) paper "Processing a language without inflections." This is not a universal opinion, however. Morphemes that are widely considered to be inflectional affixes include the noun plural marker *men* and the verb aspect markers *le* (perfective), *guo* (experiential), and *zhe* (durative) (Li & Thompson, 1981; Packard, 2000). These morphemes mark information associated with inflection in European languages, are usually unstressed (i.e., toneless) as affixes tend to be, and meet the criterion that inflection be syntactically conditioned (Stump, 1998). For example, the durative marker *zhe*, the focus of the experiments in this paper, is obligatory in certain syntactic constructions, such as when it transforms a subordinate clause into a manner adverbial, somewhat like a gerundive clause in English (Li & Thompson, 1976; Ma, 1985).

What feeds the analytical controversy over morphemes like *zhe* is that the above listed properties are also found with clitics, syntactically free grammatical morphemes that are attached to base words phonologically rather than morphologically (i.e., base and clitic share a prosodic unit but do not form a unit treated as a whole by the surface syntax; Nespor & Vogel, 1986; Wheeldon & Lahiri, 1997). For example, examples like *the man I called up's hat* show that English possessive *'s* is a clitic, not a suffix, since *up's* is not a syntactic unit despite being a prosodic whole. By contrast, the English plural *-s* is a morphological (word-forming) suffix, since a word like *hats* is treated as a whole both prosodically and syntactically. The character-based nature of Chinese orthography provides no help in determining whether a two-character sequence like *kànzhe* ("watching") should be analyzed as a clitic group or as a morphological word. Forms like *kànzhe* are thus highly parsable: neither the stem nor the inflectional markers are modified phonologically or orthographically, unlike, say, the English regular plural (see Hay, 2002, for more on the processing consequences of parsability). It is this high parsability that makes Chinese a particularly intriguing case in the search for surface frequency effects in regular inflection: *zhe* forms are about the most unlikely candidates for whole-word storage that could be imagined.

Of the criteria proposed by Zwicky and Pullum (1983) for distinguishing affixes from clitics, the one most clearly suggesting that *zhe* is an inflectional affix is selectivity with respect to its host. *Zhe* only appears with bases of the appropriate syntactic and semantic type, namely verbs describing an open-ended process or situation (Li & Thompson, 1981; Ma, 1985), while verbs that disfavor the use of *zhe* include those describing a change of state, the start or end of an action, and so-called resultative compounds (e.g., *zhǎodào* "find", literally "search-arrive"). Syntactic and semantic restrictions are not necessary with clitics (recall the English possessive), but are typical of inflection.

While semantic restrictions suggest that it is valid to classify *zhe* as an affix, they make it harder to study surface frequency effects in *zhe* forms, since such effects will tend to be confounded with semantic acceptability. That is, it seems reasonable to expect that Chinese speakers will tend to avoid using *zhe* forms with base verbs that do not describe a clearly open-ended process or situation, thereby reducing their frequency, while frequency may be used by language learners as evidence for the semantic restrictions on *zhe*; adults may also take familiarity into consideration when giving acceptability judgments, further increasing the correlation. For the forty-five *zhe* forms used in the experiments in this paper, there is indeed a significant correlation between the (log) surface frequency and acceptability (measured by pretest as described below) ($r(43) = .49, p < .001$). We should note that this kind of problem is not restricted to Chinese. Although for regularly inflected past tense verbs in English reported in Ullman (1999) failed to find any effect of surface frequency on acceptability judgments (in contrast to irregular forms, which showed a positive correlation), other judgment tasks have found surface frequency effects for regular forms in Dutch, including subjective frequency estimation tasks conducted on regular plural forms (Schreuder & Baayen, 1997).

Considerations like these led Taft (1979, 2004) to propose that apparent surface frequency effects are actually reflections of the ease with which bases and affixes are recombined after obligatory decomposition, which in turn depends on acceptability. Unlike surface frequency, which records past experience dealing with base-affix combinations (whether as wholes or as the output of a composition process), the acceptability of a base-affix combination can be computed online solely from what Taft calls the functional information associated with the base morpheme, along with knowledge of the affix's semantic restrictions. For example, Taft points out that the lower surface frequency of *seeming* relative to *growing* correlates with a semantic difference, since *seem* is stative (conflicting with the progressive aspect of *ing*) while *grow* is dynamic (compatible with progressive aspect). (Note that this posited role for functional information is quite different from the semantic neighborhood effects on inflection studied by Baayen & Moscoso del Prado Martin, in press, and Ramscar, 2002.)

Thus it is crucial to distinguish true surface frequency effects from mere acceptability effects. It is also crucial, we believe, to recognize that despite their natural correlation, these notions are logically distinct: surface frequency is an objective distributional property, while semantic acceptability depends on native-speaker knowledge. Whatever Chinese speakers use to acquire their knowledge of aspect marker semantics, token frequency is likely to be only one factor among many (note that the correlation coefficient cited above implies that variance in surface frequency in our materials explains only about 24% of the variance in acceptability). In principle, the logical distinction between surface frequency and acceptability works both ways. Frequency can vary without necessarily affecting acceptability: if a new stative verb were to enter Chinese, the first time it underwent *zhe* suffixation the resulting form would be extremely rare (only one token) yet it would still be acceptable since it obeys the semantic constraints on *zhe*. Likewise, acceptability can vary without necessarily being reflected in frequency. Corpus linguists (e.g., Manning, 2003) and sociolinguists (e.g., Labov, 1996) have often noted that speakers (or writers) can use structures that they themselves will reject as unacceptable if explicitly asked; Gibson, Schütze, and Salomon (1996) report variation in syntactic acceptability that depends on structural properties rather than frequency. In a sense, as Gibson et al. (1996) themselves note, this is just a simple corollary of the difference between production (reflected in corpus-based frequency measures) and comprehension (reflected in metalinguistic judgments), but at least in the case of *zhe*, it may also relate to the speaker's difficulty in satisfying competing constraints. *Zhe* has not only a semantic function (durative) but also a discourse function (backgrounding information; Li & Thompson, 1981). Thus a speaker may have to produce a somewhat less acceptable *zhe* form in order to achieve a larger discourse goal.

One would still like to know how speakers acquire their acceptability judgments, of course, but the two other simple possible sources we have considered fare just as poorly as surface frequency. Thus in our materials, correlations between base frequency and acceptability are entirely nonexistent (r = -.03), even when surface frequency is partialed out (r = -.07) or when only the lowest-surface-frequency items are examined (r = .07). At first glance another distributional property, competition between morphemes or allomorphs, seems to hold some promise. Surface frequency effects are more readily found for a regularly inflected word if there is an alternative irregularly inflected form (e.g., *dreamed* and *dreamt*; Ullman, 1999; Pinker, 1999) or if the affix has a homonym (Bertram et al., 2000). In such cases, the word recognition system apparently must check the base-affix combination to disambiguate the target from the competitor, and given the view of Taft (1979, 2004), this checking process may be expected to bring acceptability into play. Moreover, like acceptability, morphological competition is in principle independent of frequency. Thus it is possible to control both base and surface frequency in a pair of words [$base_1$-$affix_1$] and [$base_2$-$affix_1$], but still allow the base of one word to have an alternative affixed form [$base_1$-$affix_2$], the token frequency of which matches the token frequency of the unaffixed surface form [$base_2$] of the other. We might therefore expect the acceptability of [$base_1$-$affix_1$] to be higher than that of [$base_2$-$affix_1$], since $base_1$ appears more often in an affixing context than $base_2$.

Unfortunately, as with surface and base frequency, this attempt to reduce acceptability to distributional statistics faces serious challenges. The most likely competitor of *zhe* is another durative marker, *zài*, but unlike the cases studied in English and Dutch, it is not a homophone (indeed, since it is full-toned and thus stressed, it cannot even be cliticized), nor is it an allomorph, but a distinct morpheme. *Zài* differs from *zhe* both in syntactic behavior (it appears before the verb, not after it) and in semantic function; Smith (1991) calls it a progressive marker, as opposed to the stative imperfective marker *zhe*, with Li and Thompson (1981, p. 221) citing the minimal pair *Tā chuān-zhe píxié* "S/he is wearing his/her leather shoes" (stative imperfective) vs. *Tā zài chuān píxié* "S/he is putting on his/her leather shoes" (progressive). The same point holds for the only other verbal grammatical morphemes in Chinese, the perfective *le* and experiential *guo*, which are likewise neither homonymous nor synonymous with *zhe*. Moreover, all verb bases in our experimental materials allow all four aspect markers (*zhe*, *zài*, *le*, *guo*). Thus morphological competition, if it exists at all, cannot be all-or-none. In our materials there are negative (though not significant) correlations of (log) *zài* + verb frequency with verb + *zhe* frequency (*r* = -.22) and with *zhe* form acceptability (*r* = -.21), suggesting the possibility of weak competition between these two morphemes. However, there is no evidence of competition between *zhe* and the other two post-verbal morphemes; the correlations of verb + *le* frequency and verb + *guo* frequency are even smaller, and not consistently negative. This suggests that the competition between *zhe* and *zài*, if it is real at all, is semantic or discourse-based, with speakers choosing one or the other depending on the message to be conveyed. Finally, a regression with acceptability as independent variable and the (log) surface frequencies of *zhe*, *zài*, *le*, and *guo* forms as predictors did nothing more than reconfirm the correlation between *zhe* form frequency and acceptability discussed above. In particular, there was no evidence of any interactions between surface frequency for *zhe* and that for the other three aspect markers, as one would expect if morphological competition is what creates acceptability variation. All of the above null results also hold specifically of the materials in our final pair of experiments, where base and surface frequency were controlled and acceptability varied.

We hope that the lesson of all this is not overly dramatic: we simply do not yet know how Chinese speakers acquire their acceptability judgments. Nevertheless, judgment scores have face validity: *zhe* forms matched on surface frequency, such as *túmǒzhe* "smearing" (higher acceptability) and *zànměizhe* "praising" (lower acceptability), differ semantically in the expected way, since smearing is more likely to involve an open-ended process than is praising (see the materials for Experiments 4 and 5 in the appendix for further examples). Evidence that acceptability also has behavioral effects distinct from those of surface frequency will be highlighted when we discuss our experimental results.

Though we question Taft's (2004) reduction of acceptability to distributional statistics like surface frequency, we adopt his position that surface frequency effects are like acceptability effects in that they emerge in a late stage where morphological components are combined into a word, not during initial contact with the lexicon. For Chinese readers we do not see how it could be otherwise, given the extreme parsability of inflectional affixes and the more general fact that Chinese readers compose, not decompose, words. Thus like Taft, we predict that in a lexical decision task, surface frequency effects will only be found when implementation of the base-affix combination process helps make reliable lexical decisions. The manipulation used by Taft (2004) was to vary the nature of the nonword base-affix foils: in one condition the bases were themselves nonwords (e.g., *milphs*), so that lexical decisions did not require composition, and in the other condition the bases were real words that never take the given affixes (e.g., *mirths*); this forced participants to take base-affix combinability into account. While Taft used this manipulation to study base frequency effects, not the focus of the present paper, we adopt this nonword manipulation for our purposes, as well as another also expected to lead readers to treat *zhe* forms as wholes: greater variation in acceptability across real *zhe* forms.

We began to suspect that surface frequency effects in Chinese inflection may depend partly on acceptability after consideration of the weaknesses of Huang (2001), the first (and thus far only) study on the processing of Chinese inflection. Huang was aware of the potential influence of acceptability on the processing of *zhe* forms, and so included it as a factor in the design of her visual lexical decision experiments in addition to base and surface frequencies. The design involved pairs of experiments with stimuli matched in acceptability and base frequency but varying in surface frequency across conditions. Following previous inflection experiments, the nonword foils had nonword bases (specifically, strings of two real characters combined in ways that did not form real words or grammatical phrases). As in the experiments in the present paper, lexical verb bases were compounds (a more typical word type in Chinese than monomorphemic words). Specifically, they were verb + verb compounds, a type that is perhaps less typical in Chinese than verb + noun compounds, in order to ensure that the *zhe* marker could only appear in final position; by contrast, verb + noun compounds behave in a phrase-like fashion (Zhou, Ostrin, & Tyler, 1993), allowing aspect markers to infix after the verb morpheme. When *zhe* forms were tested in Huang's experiments, there were significant facilitative effects of surface frequency, despite the matching in base frequency and acceptability and the lack of any difference across conditions in response time when the base forms were presented in isolation.

Unfortunately, the acceptability scores used by Huang (2001) were based on a pretest with only a four-point scale, which may not have been sensitive enough to pick up subtle differences in acceptability. It is possible, therefore, that the apparent surface frequency effect was actually an effect of acceptability, just as Taft (2004) would claim. Indeed, when we conducted an improved acceptability pretest (described in the method section for our first experiment), we found that the mean acceptability score for higher-frequency *zhe* forms in Huang's experiments was significantly higher than that for lower-frequency *zhe* forms.

There were also problems with Huang's frequency estimates, which came from the Taiwanese version of Yahoo! (tw.yahoo.com). As shown by Blair, Urland, and Ma (2002), counting page hits using an Internet search engine can provide frequency estimates as accurate as the absolute number of words in a traditional corpus, and in this case there was no alternative, given the relatively low frequency of *zhe* forms. Many acceptable *zhe* forms are not found at all in the largest available corpus of Chinese used in Taiwan, the Academia Sinica Balanced Corpus of Modern Chinese (Sinica Corpus; Chen, Huang, Chang, & Hsu, 1996). Yet Yahoo! itself is much smaller than Google (www.google.com), claimed to cover over 8 billion pages in total (as of July, 2005; unfortunately it is impossible to determine exactly how many of these pages are in the traditional Chinese characters used in Taiwan). Assuming that Google estimates were more accurate, we recalculated all frequencies for Huang's materials and found that base frequencies in her experiments were not controlled either (higher-frequency *zhe* forms had significantly higher-frequency bases). Thus all frequencies reported in the present paper were estimated using Google (data collected 6 am GMT, July 14, 2005, after all of the experiments had been run). Further discussion of our frequency estimates is given in the method section for our first experiment.

Another potential problem with Huang's experiments was her failure to consider character frequency, known to play an important role in lexical decision tasks in Chinese (see reviews in Myers, in press; Taft et al., 1999). We therefore take this factor into account in our experiments. Of course there are an unlimited number of other nuisance factors to consider in lexical research; some that, for practical reasons, we did not control include syllable frequency and number of homophones, semantic ambiguity, and the semantics and pronunciation of subcharacter components (the so-called semantic and phonetic radicals). It seems probable that most of these were matched automatically in the sampling, and in fact this turned out to be the case for the frequency of semantic radicals and the number of strokes (a measure of the visual complexity of characters). More importantly, since our designs involved experimental manipulation of the context in which the base compounds were presented (in particular, the presence or absence of *zhe*), they acted as their own controls.

Due to the interpretation problems posed by nuisance factors, not to mention the tendency for surface frequency to correlate with base frequency and acceptability, we went beyond the factorial designs of our experiments (controlling all but one of the three key factors) to conduct regression analyses as well. These test or extend the same research hypotheses as our factorial analyses by using more information, in particular actual frequencies and acceptability scores rather than high vs. low categories (note that analysis of variance can be seen as a special case of the more general linear regression approach; see, e.g., Kirk, 1995). Regression analyses have the additional advantage of being able to mimic an analysis of covariance, allowing us to factor out continuous variables like character frequency. We will therefore treat the regressions with continuous predictors as equally important as the factorial analyses using categorical predictors, if not more so.

We are now ready to describe our experiments. The central goal of this paper is to look for evidence of surface (*zhe* form) frequency effects in Chinese inflection, following the general procedures of Sereno and Jongman (1997) and other previous studies. Our first experiment, then, involves varying only surface *zhe* frequency while controlling base frequency and acceptability, with the materials presented as uninflected base forms or as suffixed *zhe* forms. We expect to find a difference across conditions when surface forms are presented, since surface frequency varies, but not when base forms are presented, since base frequency is matched. All of the other experiments reported in this paper are variants on this one.

Experiment 1: Vary surface frequency only

Method

Materials.

As noted in the introduction, frequency estimates were based on Google Web page hits (collected automatically using Query Google, an online Java tool; Hayes & Ma, 2005). Due to the nature of Chinese orthography, these estimates actually count character strings, not words per se. As a check of the reliability of our estimates, we compared Web search and Sinica Corpus counts for all 25,037 two-character nouns in the Sinica Corpus (collected as part of another study). Log frequencies from the two sources were well correlated (r = .71, r^2 = .51), with the uncaptured variance most likely due to inadequacies of the Sinica Corpus; the texts comprising it are written in a more formal style than many texts on the Web, a nontrivial portion of which are written by, or aimed at, college students, and are thus more representative of the language used by our experimental participants. The Web-derived character and word frequencies were also sufficiently accurate to derive mutual information scores that clearly distinguished between words and nonwords in all of the experimental materials used in this paper (mutual information is a measure of collocation common in corpus linguistics, defined here as the base-two log of the ratio between compound frequency and the product of the two character frequencies; Church & Hanks, 1990).

Twenty-four bimorphemic verbal (verb + verb) compounds were chosen. Twelve verbs (such as *bàochí* "hold a belief") were associated with higher-frequency *zhe* forms, with a median token frequency of 21,200 (range 8,030 - 151,000), and twelve verbs (such as *jiāozhī* "interweave") were associated with lower-frequency *zhe* forms, with a median token frequency of 5,205 (range 617 - 7,650). Since the ranges did not overlap, the mean log surface frequencies (4.35 and 3.56) were significantly different ($t(22) = 5.25, p < .0001$). The two sets of verbs were matched in base frequency, that is, in how often the two-character string composing the verb stem appeared in the corpus regardless of context (median 131,000 tokens, range 42,700 - 315,000, vs. median 122,000 tokens, range 56,800 - 464,000); there was no significant difference between the mean log frequencies (5.12 vs. 5.16).

The two sets of verbs also did not differ significantly in the frequency of the first character or second character, the number of strokes in the first or second character (a measure of visual complexity), the frequency of their semantic radicals (using data from Li, Li, & Tseng, 1997), or semantic transparency, based on judgments on a seven-point scale averaged across twenty native speakers who did not participate in any of the experiments or other pretests. Because they were matched in base frequency and character frequency, the two word types also matched in mutual information, showing that there was no difference in the probability of characters appearing in these words rather than elsewhere.

To determine the acceptability of the base-affix combinations, we asked twenty university students to judge the acceptability of 123 *zhe* forms (including those used by Huang, 2001), 43 other *zhe* forms deemed acceptable by the native-speaking coauthors, 10 *zhe* forms involving verb types that should strongly disfavor the use of *zhe*, and the 5 *zhe* forms with the worst acceptability scores from the pretest reported by Huang. Judgments were made on a seven-point scale, from 1 for *mistake, cannot understand* (*cuòwù bù néng lǐjiě*) to 7 for *very correct and easy to understand* (*hěn zhèngquè qiě róngyì lǐjiě*). There was a strong association between items and scores across individual judgments, as measured by omega-squared: $\omega^2 = .19$ ($\omega^2 > .138$ implies a large effect; see, e.g., Kirk, 1995). Scores were then averaged across all judges for each item. Using these scores, the two sets of real word bases did not differ significantly in acceptability for their associated *zhe* forms: higher-frequency *zhe* forms had a mean acceptability score of 5.63 while lower-frequency *zhe* forms had a mean acceptability score of 5.58, not a significant difference.

Of relevance to the regression analyses is the fact that the acceptability scores, log base frequencies, and log surface frequencies were not collinear: the tolerances for all of the variables (i.e., $1\text{-}R^2$ with the given variable as dependent and the others as independent) were quite high (.59, .72, .78, respectively).

These real word items were complemented by a set of 24 two-character nonword foils, created by selecting characters arbitrarily from a Chinese dictionary and combining them in ways that did not create real words or grammatical phrases; foils like these are virtually ubiquitous in research on Chinese morphological processing since unlike foils composed of nonsense characters, foils composed of real characters require some degree of word-level processing to distinguish them from real words. Real words and nonwords were matched in the number of strokes in the first and second characters, and semantic radical frequency; first character frequency was also matched, but the second character of the nonwords was higher than that of the real words (the materials were originally designed using frequency estimates collected prior to the new Google estimates collected in July, 2005). The complete set of materials is given in the appendix.

One group of participants were presented only two-character strings (bases) while another group saw the same strings suffixed with *zhe* (inflected forms).

Procedure.

Participants performed a visual lexical decision task, controlled by E-Prime (Version 1.0; Schneider, Eschman, & Zuccolotto, 2002) running in Windows 98 on IBM-compatible personal computers in a sound attenuated room. In each trial, a fixation point ("+") first appeared in the center of the screen for 1 sec, which was then replaced by a horizontal (left to right) string of black 1.5-cm-high characters on a white background. All experimental items (words and nonwords) were presented in random order; filler items were not used. Responses were measured by means of buttons labeled *zhēncí* ("real word") and *fēicí* ("nonword") on (respectively) the right and left sides of a Psychology Software Tools serial response box; participants were asked to press these buttons as quickly and accurately as possible. If a participant failed to respond within three seconds, no time was recorded and the next trial began. After a practice session with 8 items, the experimental session ran with all items presented in random order. The experiment took less than 10 minutes per participant.

Participants.

Forty university students in southern Taiwan were paid for their participation. All were native speakers of Mandarin with normal or corrected-to-normal visual acuity. Twenty were arbitrarily assigned to the base group and twenty to the inflected form group.

Results

Following Sereno and Jongman (1997), we first analyzed results for the base and inflected form groups separately, then analyzed all results together. We give effect sizes in terms of 95% confidence intervals (*CI*) for the differences between the by-participant means (Loftus & Masson, 1994; Masson & Loftus, 2003); a difference in means (Δ) that is larger than the confidence interval is significant by participant (reported differences may not exactly match reported means due to rounding). Our factorial (categorical predictor) analyses involve comparing unmatched sets of items (words vs. nonwords, words with high-frequency *zhe* forms vs. words with low-frequency *zhe* forms), so significance of main effects and interactions was tested by computing *minF'* from the by-participant F_1 and by-item F_2 (Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999), with alpha set at .05. In the text below, a claim of significance without further elaboration is based on this test; the relevant statistics are shown in Table 1.

[INSERT TABLE 1 ABOUT HERE]

Bases.

For the group of participants presented with bases, analyses of variance were run by participants (within-group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or (following Sereno & Jongman, 1997) reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 83 observations (8.6%) for the reaction time analyses.

Reaction times were significantly shorter for real words (728 ms) than for nonwords (940 ms) (Δ = 212 ms, *CI* = 78 ms). Accuracy rates were also significantly higher for real words (97.9%) than for nonwords (93.8%) by participant and by item, but not by *minF'* (Δ = 4.2%, *CI* = 3.1%).

The mean reaction time for high-*zhe* base forms (739 ms) was higher than that for low-*zhe* base forms (719 ms), a difference that was marginal (i.e., $.1 > p > .05$) by participant and not significant at all by item. Mean accuracy rates (97.5% and 98.3%, respectively) were not significantly different by participant or by item.

In an attempt to confirm the marginal by-participant effect on reaction time, we conducted repeated-measures regression analyses using the simplest form of the procedure recommended in Lorch and Myers (1990). Separate regressions were calculated for each participant, with acceptability, log base frequency, and log surface frequency as predictors and reaction time as dependent variable. The resulting regression coefficients were then submitted to two-tailed one-group *t* tests to determine if the means were significantly different from zero (this procedure does not provide a value for R^2, but the degree of model fit is not relevant here). Although in this type of analysis only participants are explicitly treated as a random effect, item variability is controlled due to the inclusion of nuisance covariates as predictors (here, acceptability and surface frequency, since target items were isolated bases). The regression revealed no effects on reaction time of surface frequency or acceptability, but there was a significant effect of base frequency ($\beta = -0.10$, $t(19) = -4.45$, $p < .001$); the negative coefficient shows that the effect went in the expected direction (i.e., higher base frequency meant shorter reaction time). When we added log character frequency to the regression model, this had no effect on the other factors, and the character frequencies themselves had no significant effects.

Inflected forms.

For the group of participants presented with inflected forms, analyses of variance were run by participants (within-group) and items (between-group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 82 observations (8.5%) for the reaction time analyses.

Reaction times were significantly shorter for real words (698 ms) than for nonwords (938 ms) ($\Delta = 240$ ms, $CI = 132$ ms). Accuracy rates were also significantly higher for real words (98.4%) than for nonwords (92.1%) ($\Delta = 6.3\%$, $CI = 5.0\%$).

The mean reaction time for higher-frequency *zhe* forms (701 ms) was very close to that for lower-frequency *zhe* forms (697 ms), not a significant difference by either participant or by item. Accuracy rates (98.8% and 97.9%, respectively) were likewise not significantly different.

To examine the independent contributions of acceptability, log base frequency, and log surface frequency to reaction time, we again conducted repeated-measures regression analyses. Base frequency had a significant (and facilitative) effect on reaction time ($\beta = -0.17$, $t(19) = -3.10$, $p < .01$), while neither surface frequency nor acceptability had significant effects; see Table 2. Adding first character log frequency and second character log frequency did not change the significance of the other factors, and the character frequencies themselves did not have significant effects.

[INSERT TABLE 2 ABOUT HERE]

Combined results.

Two-way analyses of variance (morphological form [base vs. inflected form] × surface frequency [high vs. low]) were conducted on reaction times and on accuracy rates for the real words, both by participants and by items. For both reaction times and accuracy rates, no main effects were found for either factor, whether by participants or by items, and the morphological form × surface frequency interaction was also nonsignificant. These null results are summarized, along with all analyses for combined results involving Experiment 1, in Table 3.

[INSERT TABLE 3 ABOUT HERE]

Discussion

The surface frequency effects reported in Huang (2001) disappeared with the materials used in Experiment 1, as shown by the lack of a main effect in the categorical predictor analysis and in the regression with continuous predictors, and the lack of an experiment × condition interaction.

Importantly, the regression analyses revealed a facilitative base frequency effect, whether or not *zhe* was suffixed (base character frequency played no role). Due to the design of Experiment 1, base frequency varied much less than surface frequency (*SD* 0.28 vs. 0.54, $F(23, 23) = 3.78$, $p < .01$), but there was apparently still sufficient variability in base frequency for the regression to reveal effects on reaction time. This finding serves as a reminder that fundamental lexical variables can rarely be fully controlled, and the remaining variability may be sufficient to affect the results.

The lack of a surface frequency effect implies that participants apparently either did not take *zhe* into consideration at all when making their decisions, or else they treated the three-character strings as they would a freshly minted syntactic construction; this made reference to prior experience with the verb + *zhe* combinations unnecessary. By contrast, the base frequency effect shows that they processed the base verb, checking whether it contained a lexically attested combination of characters, and then made their decisions. This algorithm was available to them because the nonword foils contained nonword bases; the presence or absence of *zhe* had no effect on lexical status.

This pattern is different from what Sereno and Jongman (1997) found in their study on English, where surface frequency effects were found in experiments with a very similar design (we also failed to replicate their reverse surface frequency effect on bases). Our results suggest that the greater degree of parsability of inflection in Chinese orthography relative to English orthography may have some effect on how inflection is processed by readers: Chinese readers can fail to process base-affix combinations under certain conditions, while under apparently similar conditions, English readers cannot. Note our emphasis on reading; our experiments cannot demonstrate a difference in how inflection is processed in spoken Chinese and English, an issue we return to briefly in the general discussion.

We next ran the natural complement of these experiments, varying base frequency and controlling surface frequency (and acceptability). This time we expected the base frequency effect to emerge in the categorical predictor analysis as well.

Experiment 2: Vary base frequency only

Method

Materials.

Twenty-four bimorphemic (verb+ verb) verbal compounds were chosen. Twelve verbs (such as *yǐncáng* "hide") had higher-frequency base forms (i.e., they were themselves of higher frequency), with a median token frequency of 456,000 (range 297,000 - 1,160,000), and twelve verbs (such as *jiázá* "mingle") had lower-frequency base forms, with a median token frequency of 70,250 (range 33,600- 103,000); since the ranges did not overlap, the mean log frequencies (respectively 5.71 and 4.83) were significantly different ($t(22) = 12.06$, $p < .0001$).

The two sets of verbs were matched in the surface frequencies of the associated *zhe* forms (median 5,705 tokens, range 220 - 55,400, vs. median 7,050 tokens, range 176 - 36,600); mean log surface frequencies (3.53 vs. 3.64) were not significantly different. They were also matched in acceptability scores for the associated *zhe* forms (5.096 vs. 5.088). The two sets of verbs also did not differ in semantic transparency or the frequency, number of strokes, or frequency of the semantic radical of the first or second character. Since we varied base frequency while keeping character frequency constant, the two word types necessarily differed in mutual information: characters in low-base-frequency words appeared more often elsewhere than characters in high-base-frequency words.

The acceptability scores for the associated *zhe* forms, log base frequencies, and log surface frequencies for the associated *zhe* forms were not collinear, as shown by their high tolerances (.69, .99, .69, respectively).

These real word items were complemented by 24 two-character nonword foils identical (except for three items) to those used in Experiment 1 (i.e., nonsense bases). Real words and nonwords were matched in the frequency, number of strokes, and radical frequencies of both first and second characters. The complete set of materials is given in the appendix.

One group of participants were presented only two-character strings (bases) while another group saw the same strings suffixed with *zhe* (inflected forms).

Procedure.

The procedure was identical to that used for Experiment 1, except that input to E-Prime was made using the keyboard rather than a button box, with the "p" key on the right side of the keyboard labeled *zhēncí* ("real word") and the "q" key on the left side of the keyboard labeled *fēicí* ("nonword").

Participants.

Forty new participants from the same pool as for the previous experiment were paid for their participation.

Results

We first analyzed results for the base and inflected form groups separately, then analyzed all results together. Statistics were conducted as for Experiment 1 and are summarized in Table 4.

[INSERT TABLE 4 ABOUT HERE]

Bases.

For the group presented with bases, analyses of variance were run by participants (within group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 128 observations (13.3%) for the reaction time analyses.

Reaction times were significantly shorter for real words (636 ms) than for nonwords (760 ms) (Δ = 124 ms, *CI* = 56 ms). Accuracy rates were higher for real words (94.2%) than for nonwords (86.5%), significant by participant and by item but only marginal (i.e., .1 > p > .05) by *minF'*.

The mean reaction time for higher frequency base forms (609 ms) was significantly lower than that for lower frequency base forms (666 ms) (Δ = 124 ms, *CI* = 56 ms). The mean accuracy rate for higher frequency base forms (98.8%) was also significantly higher than that for lower frequency base forms (89.5%) (Δ = 7.7%, *CI* = 7.3%).

To look for possible effects of character frequency, we ran repeated-measures regressions with reaction time as dependent variable. When log base frequency was the sole predictor, its effect was significant (β = -0.27, $t(19)$ = -9.80, p < .0001); this pattern remained when the log frequencies of the first and second characters were added, both of which had only marginally significant effects (first character: β = -0.06, $t(19)$ = -2.01, p = .059; second character: β = -0.07, $t(19)$ = -1.77, p = .09).

Inflected forms.

For the group presented with inflected forms, analyses of variance were run by participants (between group) and items (within group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 101 observations (10.5%) for the reaction time analyses.

Reaction times were significantly shorter for real words (670 ms) than for nonwords (800 ms) (Δ = 131 ms, *CI* = 43 ms). Accuracy rates were also significantly higher for real words (96.0%) than for nonwords (89.2%) (Δ = 6.9%, *CI* = 4.8%).

In contrast to Experiment 1, there was a significant effect of condition on reaction time. *Zhe* forms containing higher frequency bases were responded to significantly faster (639 ms) than those containing lower frequency bases (702 ms) (Δ = 62 ms, *CI* = 29 ms). The mean accuracy rate for *zhe* forms with higher frequency bases (98.3%) was also significantly higher than that for *zhe* forms with lower frequency bases (93.8%) by both participant and item, but not by *minF'* (Δ = 4.6%, *CI* = 3.7%).

To examine the independent contributions of acceptability, base frequency, and surface frequency to the reaction times, we conducted repeated-measures regression analyses. We found that both log base frequency (β = -0.26, $t(19)$ = -4.72, $p < .001$) and log surface frequency (β = -0.16, $t(19)$ = -4.52, $p < .001$) had significant independent facilitative effects on reaction time, while acceptability failed to show a significant effect; see Table 5. When we added the log frequencies of the first and second characters, the pattern of results remained the same; second character frequency also had a significant facilitative effect (β = -0.12, $t(19)$ = -3.75, $p < .01$), while first character frequency did not.

[INSERT TABLE 5 ABOUT HERE]

Combined results.

Two-way analyses of variance (morphological form [base vs. inflected form] × base frequency [high vs. low]) were conducted on reaction times and on accuracy rates for the real words, both by participants and by items. The only results significant by *minF'* were main effects of base frequency for both reaction time and accuracy: forms containing higher frequency bases were responded to faster (624 ms) and more accurately (98.5%) than those containing lower frequency bases (683 ms, 91.7%) (reaction time: Δ = 60 ms, *CI* = 16 ms; accuracy: Δ = 6.8%, *CI* = 2.9%). The mean reaction time for the base group (638 ms) was lower than that for the inflected form group (670 ms), but this was only significant by item. Neither reaction time nor frequency showed any interaction. These results are summarized in Table 6.

[INSERT TABLE 6 ABOUT HERE]

Discussion

This experiment confirmed a base frequency effect, which was equally strong whether or not *zhe* was suffixed. This result was never really in doubt given the improbability of Chinese readers accessing inflected forms as wholes. However, the regression analyses also revealed, perhaps surprisingly, an independent facilitative effect of surface frequency on reaction time for *zhe* forms, without there also being an effect of acceptability. This implies that at some point before giving their responses, participants did indeed consult prior experience combining these verbs with *zhe*.

Why did this pattern emerge here, with varying base frequencies and matched surface frequency, and not in Experiment 1, with varying surface frequency and matched base frequencies? A plausible answer follows directly from the recognition that in both experiments, bases were activated before *zhe* forms. Due to their designs, base frequency of course varied much less in Experiment 1 (*SD* 0.28) than in Experiment 2 (*SD* 0.48) ($F(23, 23) = 2.97$, $p < .01$). Thus decisions for the inflected form group in Experiment 1 could reliably be made by considering the bases alone; there was no need to check the verb + *zhe* combination to distinguish them from strings containing truly nonword bases. By contrast, in Experiment 2 the greater variability in base familiarity made the lexical status of some bases less certain, and so the familiarity of the base-affix combinations was checked as supplementary assistance in the decision-making process.

The fact that combinability was key and not whole-word storage is demonstrated by another finding from Experiment 2: whereas character frequency had at best a marginal effect on reaction time when bases were presented in isolation, there was a significant facilitative effect of character frequency on reaction time when *zhe* was added, but only for the second character. We posit that two factors conspired to encourage participants to give particular consideration to the second character. First, uncertainty over the lexical status of the base forms led them to consider combinability with *zhe*, and the second character of the base is the one immediately adjacent to *zhe*. Second, since our bases were all verb + verb compounds, the combination of the second verb and *zhe* was a potential word in its own right. These considerations imply that there is no contradiction between Chinese readers accessing words via characters (even those composing the bases) and the emergence of surface frequency effects in inflected forms.

Though the appearance of a surface frequency effect suggests that memory traces for whole *zhe* forms affect word access, there is still a need for further confirmation from an experiment explicitly designed to reveal such an effect. One way to force readers to take base-affix combinability into account, which is what we suggest gives rise to surface frequency effects, is to adopt the manipulation used by Taft (2004) and change the nonword foils so that the presence or absence of *zhe* makes a difference to the determination of lexical status. That is, the bases of the foils should be real words, but chosen so that their combination with *zhe* is illegal; we did this by using noun bases, which never allow the suffixation of verbal aspect markers like *zhe*. For the real word targets we went back to the materials of Experiment 1, which were matched on base frequency and acceptability but varied in surface frequency. Because the bases in this experiment were all real words, there was only one group of participants, who were presented with inflected forms.

Experiment 3: Vary surface frequency only, noun + *zhe* foils

Method

Materials.

The same real *zhe* forms as in Experiment 1 were used. Nonword foils were two-character nominal compounds followed by *zhe*. These noun bases were matched with the verb bases of the real word items in the frequencies, number of strokes, and frequency of the semantic radicals in both the first and second characters, as well as in word frequency. The complete set of materials is given in the appendix.

Procedure.

The procedure was identical to that used for Experiment 2.

Participants.

Twenty new participants from the same pool as for the previous experiments were paid for their participation.

Results

Analyses of variance were run by participants (within group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 90 observations (9.4%) for the reaction time analyses. *MinF'* statistics were computed as for Experiment 1 and are summarized in Table 7.

[INSERT TABLE 7 ABOUT HERE]

Reaction times were significantly shorter for real words (669 ms) than for nonwords (710 ms) (Δ = 41 ms, *CI* = 17 ms). Accuracy rates for real words (95.2%) and nonwords (94.0%) were not significantly different by participant or by item.

Higher-frequency *zhe* forms were responded to more quickly (650 ms) than lower-frequency *zhe* forms (690 ms), and this difference was significant by participant but only marginally so by item (i.e., $.1 > p > .05$) and was thus not significant by *minF'* (Δ = 41 ms, *CI* = 34 ms). Accuracy rates for higher and lower frequency *zhe* forms (95.4% and 95.0%, respectively) were not significantly different by participant or by item.

Because the by-item analysis just missed significance at the .05 level, we ran a regression on the item reaction times averaged across participants using log surface frequency as the sole predictor. Since a one-factor between-groups analysis of variance is equivalent to a regression with a single dummy predictor (e.g., 1 if high surface frequency and 0 if low surface frequency), the new regression was testing the same hypothesis but with more complete information (the actual frequencies rather than the categories defined by them). The result showed a clear facilitative effect of surface frequency by item (β= -0.64, $t(22) = -3.89$, $p < .001$).

To examine more fully the independent contributions of acceptability, base frequency, and surface frequency to the reaction times, we again conducted repeated-measures regression analyses. These confirmed an independent effect on reaction time of log surface frequency (β= -0.18, $t(19) = -2.89$, $p < .01$), and found no significant effect of log base frequency. Nevertheless, acceptability also had an independent effect (β= -0.12, $t(19) = -2.72$, $p < .05$), and like surface frequency, was negatively correlated with reaction time, implying facilitation; see Table 8.

[INSERT 8 ABOUT HERE]

This pattern of results remained when we added the log frequencies of the first and second characters, but now we also found an effect of character frequency; the frequency of the second character had a significant facilitative effect on reaction time (β = -0.09, $t(19)$ = -2.20, $p < .05$), while that of the first character had no effect.

Combined results.

In order to obtain a more complete picture of the effect of the nonword manipulation, we conducted two-way analyses of variance (foil type [nonword base vs. noun base] × surface frequency [high vs. low]) on reaction times and on accuracy rates for the real words for the inflected form group in Experiment 1 vs. Experiment 3, both by participants and by items. For reaction times, the only notable result was a main effect for foil type that was significant in the by-item analysis; the mean reaction time with the inflected form group in Experiment 1 (with foils containing nonword bases) was 699 ms, while that in Experiment 3 (with foils containing noun bases) was 670 ms. However, this difference was not significant by participants, and there was no effect of surface frequency either by participants or by items. The foil type × surface frequency interaction was marginal by both participants and by items (i.e., $.1 > p > .05$): Experiment 1 showed virtually no effect of surface frequency (high vs. low surface frequency), while Experiment 3 showed a large effect. Accuracy rates showed only a main effect for foil type that was significant both by participant and by item, with accuracy for the inflected forms in Experiment 1 (98.3%) higher than for Experiment 3 (95.2%) (Δ = 3.1%, *CI* = 2.9%), but this was not significant by *minF'*. These results are summarized in Table 3.

Discussion

Using the same real word materials as Experiment 1, which had failed to find any effect of surface frequency or acceptability but only an effect of base frequency, Experiment 3 now found a facilitative effect of surface frequency (both in the categorical predictor analysis and in the regression) and an independent facilitative effect of acceptability (in the regression). The manipulation that led to these changes was the use of nonword foils that differed from real word targets only in the acceptability of *zhe* suffixation, a situation that forced readers to process the entire three-character string rather than just the first two characters forming the base. Surface frequency had an effect because of the readers' reference to prior experience making these base-affix combinations, and acceptability did as well because computation of acceptability assisted in making lexical decisions, to distinguish legal verb + *zhe* forms from illegal noun + *zhe* forms.

The appearance of surface frequency effects does not mean that readers were treating *zhe* forms as indivisible wholes. This is shown by the facilitative effect of second character frequency, just as was seen in Experiment 2 using different materials. We believe the reason for this is the same as posited in the discussion for the previous experiment: participants gave special consideration to the character immediately adjacent to *zhe*. In this experiment the appearance of this character frequency effect was accompanied by the disappearance of the base frequency effect, but we should not overinterpret this null result.

Another difference between the results for Experiments 1 and 3 seems less meaningful. The nonword reaction times for these two experiments are respectively 940 ms (*SD* 296 ms) and 710 ms (*SD* 82 ms), and the combined results for Experiments 1 and 3 show a significant difference in reaction time for real words, though only by item. Why were overall reaction times in Experiment 1 slower than those in Experiment 3? Whatever the explanation or explanations are, they must also deal with the fact that the base-only group in Experiment 1 showed similarly slow nonword responses (*M* 940 ms, *SD* 296 ms); compare those from the base-only group in Experiment 2 (*M* 760 ms, *SD* 151 ms). We also note that variance with nonwords was notably higher for the inflected form group in Experiment 1 vs. Experiment 3 (*SD* 296 ms vs. 82 ms), and that while accuracy rates for nonwords did not differ (92.1% vs. 93.9%), those for real words did (98.3% vs. 95.2%). We can only speculate that unknown differences across the groups of participants are to blame for these anomalies (Experiment 1 was conducted long before Experiments 2 and 3), but whatever the explanation, it does not seem to relate to the nonword manipulation itself.

From the main results of this experiment we concluded that surface frequency exerts its influence during the base-affix combination process. But is it truly surface frequency and not acceptability that operates at this stage? One way to address this question would be to allow acceptability to vary across real word conditions, rather than controlling it as in the previous experiments (a manipulation made possible by the logical independence of frequency and acceptability, as explained in the introduction). Such variation should trigger a processing mode in which base-affix combinability is taken into account, thereby resulting in a surface frequency effect. However, if the nonword foils contain nonword bases, acceptability by itself could not help to make the final lexicality decision and should therefore not predict reaction time.

Experiment 4: Vary acceptability only

Method

Materials.

Twenty-four bimorphemic (verb-verb) verbal compounds were chosen. Twelve verbs (such as *túmǒ* "smear") were associated with higher-acceptability *zhe* forms on the seven-point scale, with a mean acceptability score of 5.91 (*SD* 0.06, range 5.40 - 6.25), and twelve verbs (such as *zànměi* "praise") were associated with lower-acceptability *zhe* forms, with a mean acceptability score of 5.04 (*SD* 0.11, range 4.30 - 5.35); since the ranges did not overlap, these mean scores were significantly different ($t(22) = 7.06$, $p < .0001$). Note that acceptability scores never went below 4.3 on the seven-point scale, so no item could be considered ungrammatical.

The two sets of verbs were matched in base frequency (median 150,500 tokens, range 33,600 - 426,000, vs. median 190,500 tokens, range 42,700 - 464,000), with no significant difference between the mean log base frequencies (5.17 vs. 5.21). They were also matched in surface frequency (median 6,180 tokens, range 524 - 10,400, vs. median 4,285 tokens, range 607 - 12,600), with no significant difference between the mean log surface frequencies (3.59 vs. 3.50). Finally, the two sets of verbs also did not differ significantly in mutual information or pretests for semantic transparency, nor in the frequency, number of strokes, or frequency of the semantic radicals in the first and second character.

The acceptability scores of the associated *zhe* forms, log base frequencies, and log surface frequencies of the associated *zhe* forms were not collinear, as shown by their quite high tolerances (.92, .98, .90, respectively).

These real word items were complemented by 24 two-character nonword foils identical to those used in Experiment 2 (i.e., nonsense bases). Real words and nonwords were matched in frequency, number of strokes, and radical frequency for both the first and second characters. The complete set of materials is given in the appendix.

One group of participants were presented only two-character strings (bases) while another group saw the same strings suffixed with *zhe* (inflected forms).

Procedure.

The procedure was identical to that used for Experiment 2.

Participants.

Forty new participants from the same pool as for the previous experiments were paid for their participation.

Results

We first analyzed results for the base and inflected form groups separately, then analyzed all results together. *MinF'* statistics were conducted as for Experiment 1 and are summarized in Table 9.

[INSERT TABLE 9 ABOUT HERE]

Bases.

For the group presented with bases, analyses of variance were run by participants (within group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 113 observations (11.8%) for the reaction time analyses.

Reaction times were significantly shorter for real words (634 ms) than for nonwords (722 ms) (Δ = 88 ms, *CI* = 33 ms). Accuracy rates were identical for real words and nonwords (92.1%).

Reaction times for base forms associated with higher-acceptability *zhe* forms (636 ms) were very close to those associated with lower-acceptability *zhe* forms (633 ms), not significantly different by participant or item. Mean accuracy rates were also not significantly different by participant or item (91.7% vs. 92.5%).

To look for possible effects of character frequency, we ran repeated-measures regressions with reaction time as dependent variable. When log base frequency was the sole predictor, its effect was significant (β = -0.14, $t(19) = -3.95$, $p < .001$), but when log frequencies of the first and second characters were added, base frequency lost its effect and only first character frequency had a significant (facilitative) effect (β = -0.36, $t(19) = -3.48$, $p < .01$).

Inflected forms.

For the group presented with inflected forms, analyses of variance were run by participants (within group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 96 observations (10.0%) for the reaction time analyses.

Reaction times were significantly shorter for real words (652 ms) than for nonwords (770 ms) (Δ = 118 ms, *CI* = 48 ms). Accuracy rates were not significantly higher for real words (94.8%) than for nonwords (93.5%).

The difference in mean reaction time for higher-acceptability *zhe* forms (650 ms) and for lower-acceptability *zhe* forms (653 ms) was not significant by participant or by item, nor was the mean accuracy rate for higher-acceptability *zhe* forms (93.8%) significantly different from that for lower-acceptability *zhe* forms (95.8%).

To examine the independent contributions of acceptability, base frequency, and surface frequency to the reaction times, we conducted repeated-measures regression analyses. We found that both log base frequency (β= -0.21, $t(19) = -4.97$, $p < .0001$) and log surface frequency (β= -0.11, $t(19) = -3.35$, $p < .01$) had significant independent facilitative effects on reaction time, while acceptability failed to show a significant effect; see Table 10. When we added the log frequencies of the first and second characters, the pattern of results remained the same; first character frequency also had a significant facilitative effect (β= -0.08, $t(19) = -2.23$, $p < .05$), while second character frequency did not.

[INSERT TABLE 10 ABOUT HERE]

Combined results.

Two-way analyses of variance (morphological form [base vs. inflected form] × acceptability [high vs. low]) were conducted on reaction times and on accuracy rates for the real words, both by participants and by items. For both reaction times and accuracy rates, no main effects were found for either factor, whether by participants or by items, and the interaction was also nonsignificant by both participants and by items. These null results are summarized in Table 11.

[INSERT TABLE 11 ABOUT HERE]

Discussion

The regression revealed that both surface frequency and base frequency had significant facilitative effects on reaction time for *zhe* forms, but there was no evidence for a direct effect of acceptability, neither in the categorical predictor analysis nor the regression.

These results make sense within the view espoused earlier, according to which acceptability plays only an indirect role in lexical access. Even though the nonword foils used in Experiment 4 had nonword bases, there was greater variation in the acceptability of *zhe* forms across the real words compared with earlier experiments: the standard deviation of the acceptability scores for real word materials in Experiment 1 was 0.12, while that in Experiment 4 was 0.53, significantly higher ($F(23, 23) = 2.36$, $p < .05$). This variation apparently led processing resources to be devoted to base-affix combinations, resulting in the influence of surface frequency on response times. Unlike the situation in Experiment 3, however, acceptability itself was ultimately irrelevant for determining lexical status, and so there was no direct effect of acceptability on reaction time. (Note also that these results help explain what might have happened in the experiments of Huang, 2001, in which acceptability and surface frequency were unintentionally allowed to covary; her effect may have truly been a surface frequency effect, but one that was triggered by acceptability variation.) The finding that base frequency also had an independent effect on reaction time indicates that decisions were being made partly on the lexical status of bases rather than solely on *zhe* forms as wholes, consistent with the view that surface frequency effects arise after only the components have received some initial processing.

Interestingly, character frequency effects show a different pattern from the previous experiments. When *zhe* forms were displayed, both character frequency and base frequency had independent effects, unlike Experiment 3, where the base frequency effect was replaced by an effect of second character frequency. With these materials, then, participants apparently did not consider combinability with *zhe* until they had first constructed the bases, and during the base-construction process base frequency came into play. One possible reason why this happened is that variation in acceptability across the real *zhe* forms forced participants to consider the whole base, since it was this that determined acceptability with the following *zhe*, not just the second character. By contrast, in Experiment 3 acceptability across real words and nonwords was determined solely by syntactic category of the base, which for real word bases (verb-verb compounds) was reliably indicated by that of the second character.

We see, then, that acceptability variation both within words and across words and nonwords can trigger a surface frequency effect, albeit by somewhat different mechanisms. A natural question to ask is what happens when we use a combination of both types of variation. In particular, since noun + *zhe* foils make acceptability relevant in the making of lexical decisions, as we saw in Experiment 3, we expect that if we combine such nonwords with real *zhe* words varying in acceptability, acceptability should again have a significant effect on reaction time, even in the categorical predictor analysis.

Experiment 5: Vary acceptability only, noun + *zhe* foils

Method

Materials.

The same real *zhe* forms as in Experiment 4 were used. Nonword foils were the same noun + *zhe* forms used in Experiment 3. The noun bases were matched with the verb bases of the real word items in the frequencies, number of strokes, and radical frequencies of the first and second characters, as well as in word frequency. The complete set of materials is given in the appendix.

Procedure.

The procedure was identical to that used for Experiment 2.

Participants.

Twenty new participants from the same pool as for the previous experiments were paid for their participation.

Results

Analyses of variance were run by participants (within group) and items (between group) for both reaction times and accuracy rates. All reported means are from the by-participant analyses. No errors or reaction times further than 2 standard deviations from a participant's mean reaction time were included in the reaction time analyses, representing a loss of 105 observations (10.9%) for the reaction time analyses. *MinF'* statistics were conducted as for Experiment 1 and are summarized in Table 12.

[INSERT TABLE 12 ABOUT HERE]

Reaction times were significantly shorter for real words (719 ms) than for nonwords (775 ms) (*Δ* = 56 ms, *CI* = 47 ms). Accuracy rates for real words (92.7%) and nonwords

(93.5%) were not significantly different by participant or by item.

Higher-acceptability *zhe* forms were responded to more slowly (728 ms) than lower-acceptability *zhe* forms (705 ms), a difference that was marginally significant by both participants and items (i.e., .1 > *p* > .05). However, accuracy rates showed the opposite pattern, with a higher accuracy rate (96.3%) for higher-acceptability *zhe* forms than for lower-acceptability *zhe* forms (89.2%), a difference that was significant by both participant and item though (*Δ* = 7.1%, *CI* = 5.8%) not by *minF'*.

To look for any hidden contributions of acceptability, base frequency, and surface frequency to the reaction times in Experiment 4, we again conducted repeated-measures regression analyses. None of these factors had significant effects on reaction time. This remained true when we added log frequencies of the first and second characters to the model, which were themselves nonsignificant. These null results are summarized in Table 13.

[INSERT TABLE 13 ABOUT HERE]

Combined results.

In order to obtain a more complete picture of the effect of the nonword manipulation, we conducted two-way analyses of variance (foil type [nonword base vs. noun base] × acceptability [high vs. low]) on reaction times and on accuracy rates for the real words with the inflected form group in Experiment 4 and Experiment 5, both by participants and by items. For reaction times, the only notable effect was a main effect for foil type that was significant in the by-item analysis; the mean reaction time for inflected forms in Experiment 4 (with *zhe* forms containing nonword bases) was 652 ms, while that in Experiment 5 (with *zhe* forms containing noun bases) was 717 ms. However, this difference was only marginally significant by participants (i.e., .1 > *p* > .05), and there was no significant effect for acceptability either by participants or by items. The foil type × acceptability interaction was marginal by participants but not significant by items; the trend implies that only when noun-base foils were used (Experiment 5) did acceptability affect reaction times (higher-acceptability slower). For accuracy rates, the most notable result was an interaction between foil type and acceptability that was significant both by participant and by item, but not by *minF'*; accuracy rates only showed a difference relating to acceptability (higher-acceptability more accurate) with noun-base foils (Experiment 5). There was also a nonsignificant trend, both by participant and by item, for higher accuracy rates for Experiment 4 (94.8%) than Experiment 5 (92.7%) (*Δ* = 2.1%, *CI* = 5.4%). These results are summarized in Table 11.

Discussion

As predicted, the combination of varying acceptability in the real words and noun + *zhe* nonword foils produced reliable effects of acceptability in the categorical predictor analyses, using the same real words that showed no such effects with nonword-base foils. However, these effects were strikingly different in nature from any found in the previous experiments. First, the effect of acceptability on reaction time was inhibitory rather than facilitatory, with slower responses for higher-acceptability forms. Second, there was a speed-accuracy trade-off: faster responses were accompanied by lower accuracy rates. In particular, higher-acceptability words showed more accurate responses when foils had noun bases (Experiment 5), whereas there was no effect of acceptability on accuracy with the same real *zhe* forms when foils had nonword bases (Experiment 4). Finally, the regressions on reaction time using continuous predictors found no significant effects of any kind, presumably because observation-by-observation correlations were washed out by noise introduced in the speed-accuracy trade-off.

The previous experiments showed the cooccurrence of faster responses with higher accuracy rates typical of lexical decision tasks, allowing us to treat these as two measures of the same thing, namely ease of access. The speed-accuracy trade-off in the present experiment suggests that something different is going on here. One way to characterize this difference would be in terms of a model in which observers vary between responding perfectly and guessing; a task that encourages guessing will show a speed-accuracy trade-off because responses will be faster yet more error-prone. Townsend and Ashby (1983, p. 259) note, paradoxically, that a task tends to show speed-accuracy trade-offs when discriminations are easier, and suggest that this may occur because participants in an easy task become bored and so adopt the guessing strategy (p. 279). The guessing model therefore implies that Experiment 5 was easier for participants than Experiment 4, where no trade-off was observed, but it is not obvious that this is correct. The discrimination necessary in Experiment 4 involved the lexical status of bases, which required retrieval of memory traces, while that in Experiment 5 involved the legality of base-affix combinations, which in principle does not require retrieval of memory traces since it can be determined via consultation of grammatical rules (i.e., the selectional restrictions on *zhe*). Which of these tasks is harder? A hint that it was the latter, contrary to the guessing model, is given by the slower overall reaction times for Experiment 5 (significant by item and marginally so by participant), paralleled by a nonsignificant trend in overall accuracy rates (lower for Experiment 5).

We thus suggest an alternative interpretation: In contrast to previous experiments, participants in Experiment 5 were not merely deciding if items were words, but also attempting to detect violations of the selectional restrictions of *zhe*. This strategy was triggered because items varied greatly in the degree to which they violated these selectional restrictions, partially independently of their lexical status, and among real words, entirely independently of their frequency. Of course we cannot say that familiarity was entirely irrelevant. Accuracy rates, defined in terms of lexical status rather than acceptability, remained relatively high, and lexical status had the usual facilitative effect on reaction time: it took longer to reject noun + *zhe* forms than to accept verb + *zhe* forms, a pattern that should have been reversed if participants were only trying to detect violations of selection restrictions. Nevertheless, once this extreme difference in familiarity (quite familiar vs. absolutely no prior experience) had exerted its effect, we must conclude that a violation-detection strategy must have come into play in order to explain to the speed-accuracy trade-off in words, where more acceptable forms (e.g., *túmǒzhe* "smearing") were accepted after some delay whereas less acceptable forms (e.g., *zànměizhe* "praising") were quickly rejected. The latter response is not a misclassification on the hypothesis that the task had become (once familiarity had been dealt with) a violation-detection task: *zànměizhe* truly represents a greater violation of *zhe* selectional restrictions than *túmǒzhe*, and the speed of response depends on how quickly such a violation is detected. Acceptability judgments are almost always studied in offline tasks, so it is unknown how common it is for them to be associated with speed-accuracy trade-offs (the speeded acceptability judgment task in Garnham, Oakhill, & Cain, 1998, did not show such a pattern). Nevertheless, we feel that a model appealing to grammatical knowledge is, at the very least, no less plausible than one relying on the equally mysterious notion of boredom, posited by the guessing model.

Regardless of what the proper model may be for the anomalous results in Experiment 5, the fact remains that they are anomalous and relate somehow to acceptability, not frequency. Thus this experiment may have failed to shed further light on surface frequency effects, but it does confirm our prediction that the use of inflected foils with real-word bases brings acceptability into play. More importantly, it reinforces our assumption that surface frequency and acceptability do not measure the same underlying psychological factor.

General Discussion

Our experiments provide support for three key claims: (1) Chinese inflection does indeed show facilitative surface frequency effects, (2) they only appear under conditions encouraging the combination of base-affix sequences into linguistic units (presumably words), and (3) they are truly effects of frequency and not of acceptability. Taken together, these claims imply a model that ascribes surface frequency effects to a late word composition stage rather than to lexical storage of whole words, but also one in which lexical information (base-affix combination frequency) influences the composition process.

Claim (1) is based on the fact that three of the five experiments showed facilitative surface frequency effects on reaction times. All three showed this effect in the regressions using continuous predictors, and the one of these that varied surface frequency in the factorial design (Experiment 3) also showed it in the categorical predictor analysis. Additional replications of the effect include a pilot we conducted of Experiment 4 using mostly different materials, and perhaps also the experiments reported in Huang (2001).

Claim (2) is based on the observation that surface frequency effects were only found under certain circumstances, while facilitative base frequency effects were virtually ubiquitous across all eight experiments regardless of the design. This conflicts with the predictions of a model where inflected forms are stored and retrieved as wholes, according to which surface frequency effects should be an automatic consequence of making lexicality decisions on inflected forms (in any case, this model is implausible from consideration of Chinese orthography alone).

Instead, surface frequency effects only appeared when the design of the materials encouraged participants to apply a processing mode that took base-affix combinability into account. We were able to trigger this mode via three different manipulations. The first was to follow Taft (2004) in the use of foils with real word bases, which made the lexical status of the experimental items depend on base-affix combinability (Experiment 3). The second manipulation involved varying the acceptability of the lexical inflected forms themselves, which also served to make participants aware of the importance of base-affix combinability in making their decisions (Experiment 4). The third, and perhaps the most surprising, manipulation was to vary the frequency of the bases while matching surface frequency (Experiment 2). Since we believe that readers activate bases before considering the inflected forms as wholes, variation in base frequency made a base-only decision process more risky, since less-familiar bases were easier to mistake for nonwords; hence participants in the inflected form group in Experiment 2 took base-affix combinability into consideration as well.

Finally, claim (3), that surface frequency effects are not acceptability effects in disguise, follows from several pieces of converging evidence. In addition to surface frequency and acceptability being logically distinct properties of base-affix combinations, these two factors were not perfectly correlated in the materials used in our experiments; with $r = .49$, 76% of the variance in acceptability was not explained by surface frequency variance. This is just what we would expect of metalinguistic judgment scores, which must be generated via consultation of more sources than mere frequency; in the case of *zhe*, this presumably involves whatever generalizations Chinese speakers have induced about its semantic restrictions. As only partly correlated, the two factors were separable in the regression analyses, which confirmed that surface frequency could affect reaction time even with acceptability factored out.

Surface frequency and acceptability also showed quite distinct behavioral effects. As a distributional property reflecting prior experience with base-affix combinations, facilitative surface frequency effects were expected whenever bases and affixes were combined during reading, and this expectation was fulfilled, as discussed above. By contrast, acceptability only predicted reaction times when it helped in the discrimination of words and nonwords (i.e., in Experiments 3 and 5, where the nonwords had ungrammatical base-affix combinations). Elsewhere acceptability only had an indirect effect, in particular in Experiment 4, where acceptability variation among real words made it beneficial to take base-affix combinations into account (thereby causing the appearance of surface frequency effects). The fundamental difference between surface frequency and acceptability was revealed most strongly in Experiment 5, where acceptability variation across the real words and ungrammatical base-affix combinations in the nonwords induced participants to make acceptability judgments in addition to lexicality judgments, resulting in an inhibitory acceptability effect on reaction time and a speed-accuracy trade-off.

Recognizing surface frequency effects as genuine forces us to conclude that the base-affix combining process automatically makes reference to memory traces of previous instances when this process took place. That is, it is not the case that surface frequency effects are really due to readers accessing the functional information of the component morphemes and then computing acceptability on the fly. We see no contradiction between this conclusion and our rejection of whole-word access of Chinese inflected forms. In fact, storing information from past experience about the combinability of characters would seem to be an eminently practical habit for readers who face the problem of finding words in an orthographic system that does not demarcate them.

In our regression analyses we also looked at the effects of character frequency. Though we did this primarily to get a handle on a possible nuisance factor, these analyses turned out to provide further insight into how the base-affix combination process is carried out. Character frequency often had facilitative effects on reaction time, independent of surface frequency effects, though sometimes adding character frequency to the model caused base frequency effects to disappear. This is consistent with our fundamental assumption that readers looked up individual characters in memory before making taking any further step. The precise nature of the character frequency effects was not fully consistent across all experiments, not even when only bases were presented: they were entirely absent with the base-only group in Experiment 1, affected both characters but only marginally with the base-only group in Experiment 2, and were significant with the base-only group in Experiment 4, but only for the first character. Presumably this variation was due to fluctuations in the complex and interacting factors that are known to play roles in compound processing in Chinese, but we refuse to speculate further given the virtually total absence of previous research on the processing of verbal compounds (a gap also noted in the review by Myers, in press). Nevertheless, it is interesting to note that all three of the experiments showing surface frequency effects with suffixed items (Experiments 2, 3, and 4) also showed significant character frequency effects, in one case (Experiment 2) even when character frequency effects for the base-only group were merely marginal. Moreover, in two of these three experiments (Experiments 2 and 3), it was the second character alone that showed this effect, suggesting that readers were more concerned with its lexical properties than with those of the other character. It thus seems that at least in some cases, readers considered combinability between the second character and the suffix separately from the combinability of the characters in the base compound itself, an option that was available to them due to our use of verb + verb compounds.

As a methodological aside, the fact that character frequency effects were only revealed in the regressions should be seen as an argument for designing experiments on lexical access with regression specifically in mind, a point recently made with particularly forcefulness in Baayen, Tweedie, and Schreuder (2002) and Baayen (2004). Throughout this paper we have emphasized how our categorical predictor analyses threw out useful data and ignored the effects of covarying factors, nuisance or otherwise; Baayen (2004), building on Harrell (2001) and others, provides simulations demonstrating the serious consequences of such data loss. Note that the mathematical equivalence between regression and the tests more familiar in experimental research (*t* tests and analysis of variance) undermines any counterargument that regressions merely test correlation, not causation, and are therefore more appropriate for observational studies than true experiments. One could reply further that by-item analyses in lexical research do not test truly experimental hypotheses anyway, since lexical properties are inherent, not manipulated. In any event, our key finding of surface frequency effects was found, where they were expected to be found, in both regression analyses and the more familiar categorical predictor analyses.

We have made no secret in this paper of our preference for a base-first approach like that advocated in Taft (1979, 2004) over a whole-word access model like that advocated in Sereno and Jongman (1997), at least for Chinese readers. Our conclusion that Chinese readers first parse out the bases of inflected forms and only later incorporate the affix may seem to imply that we also reject the dual-route model, whereby readers toggle between whole word access and access mediated by the component morphemes, depending on the absolute and relative frequencies of the bases and surface forms (e.g., Alegre & Gordon, 1999; Baayen et al. 1997; Bertram et al. 2000; New et al. 2004). In fact, we believe that the data given in this paper are insufficient to choose between the base-first and dual-route models. At best we can say only that the dual-route model is currently incomplete; it does not incorporate the context effects induced by foil type or variability of lexical properties across real words, nor does it accommodate the special characteristics of Chinese orthography (the lack of word boundaries and conflation of orthographic and morphemic units). Like the base-first model, however, the dual-route model as formalized by Baayen and colleagues also assumes that acceptability comes into play after possible components have been identified, at a so-called licensing stage when ungrammatical parses are rejected. The dual-route model also has the advantage of being formally explicit (indeed computationally implemented). Thus it is not obvious that a suitably updated version would be unable to handle the base and surface frequency effects seen in our experiments. In particular, the lack of surface frequency effects in Experiment 1 seems to have a ready explanation in terms of ease of segmentation and licensing, just as we have been assuming.

There are a number of obvious ways in which this work can be extended. Using regressions to remove the effects of nuisance variables only works if we have information about these variables. Unfortunately, aside from word and character frequencies, Chinese as yet has no comprehensive database of many factors that have been well-studied in languages like English, so they have to be measured anew for each new study; in this study we pretested only semantic transparency and base-affix acceptability. Of course, even with well-studied languages only a small subset of such factors are considered in practice. Another way to go beyond the present study would be to examine other inflectional affixes, such as the plural suffix *men*, which is only used with [+human] nouns and thus like *zhe* has semantic restrictions that could be exploited in an acceptability manipulation. It might also be useful to replicate our study in spoken Chinese, which may help clarify if the uniquely Chinese results relate to differences in language or merely orthography (e.g., the lack of word boundary marking). Of course, such a study would also require consideration of a new set of nuisance variables (e.g., syllable frequency, cohort size, uniqueness point, neighborhood density, phonotactic probability), and on top of this, reliable frequency estimates for spoken words are very hard to come by for Chinese (spoken and written Chinese differ considerably due to remnants of classical writing traditions). Moreover, this suggestion works both ways. We already know from New et al. (2004) that with visually presented inflected forms, frequency effects are similar across English, Dutch, and French, all of which have alphabetic orthographies. What we do not know is how the situation would look in spoken word processing in these languages. We suspect that surface frequency effects will be even more robust in spoken word processing, since listeners face a word segmentation problem roughly comparable to that faced by Chinese readers (Shillcock, 1990). At least in the case of spoken Chinese compounds, surface frequency effects have been found in both of the studies where they have been looked for (Myers & Gong, 2002; Zhou & Marslen-Wilson, 1994).

In this study we have shown that frequency information about base-affix combinations in Chinese is lexically stored, though not always used. Presumably such information is stored not only for words, but also for word sequences (collocations); consistent with this, Vogel Sosa and MacFarlane (2002) found that English listeners took longer to detect the word *of* in higher-frequency collocations, as if such collocations were harder to decompose. We trust that such results should not come as particularly shocking; human memory is vast, and there is nothing preventing the storage of frequency information about any linguistic unit, regardless of how readily decomposed or composed it is. In the end, the key result of the present study can perhaps be best taken as a reductio ad absurdum to show that surface frequency effects are insufficient evidence for whole-word access of inflected forms, since such effects are even found in a "language without inflections."

References

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language, 40*, 41-61.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers, 1*, 1-45. University of Alberta, Canada.

Baayen, R. H., & Moscoso del Prado Martin, F. (in press). Semantic density and past-tense formation in three Germanic languages. *Language, 81*.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language, 37,* 94-117.

Baayen, R. H., Schreuder, R., De Jong, N., & Krott, A. (2002). Dutch inflection: The rules that prove the exception. In S. Nooteboom, F. Weerman, F. Wijnen (Eds.), *Storage and computation in the language faculty* (pp. 61-92). Dordrecht: Kluwer Academic Publishers.

Baayen, R. H., Tweedie, F. J., & Schreuder, R. (2002). The subjects as a simple random effect fallacy: Subject variability and morphological family effects in the mental lexicon. *Brain and Language*, 81, 55-65.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26* (2), 489-511.

Blair, I. V., Urland, G. R., & Ma, J. E. (2002). Using Internet search engines to estimate word frequency. *Behavior Research Methods, Instruments, & Computers, 34* (2), 286-290.

Burani, C., Salmaso, C., & Caramazza, A. (1984). Morphological structure and lexical access. *Visible Language, 18*, 342-352.

Chen, K.-J., Huang, C.-R., Chang, L.-P., & Hsu, H.-L. (1996). SINICA CORPUS: Design methodology for balanced corpora. *Language, Information and Computation, 11*, 167-176.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, 16*, 22-29.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335-359.

Garnham, A., Oakhill, J., & Cain, K. (1998). Selective retention of information about the superficial form of text: Ellipses with antecedents in main and subordinate clauses. *The Quarterly Journal of Experimental Psychology, 51A* (1), 19-39.

Gibson, E., Schutze, C. T., & Salomon, A. (1996). The relationship between the frequency and the processing complexity of linguistic structure. *Journal of Psycholinguistic Research, 25* (1), 59-92.

Harrell, Jr., F. E. (2001). *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. Berlin: Springer.

Hay, J. (2002). From speech perception to morphology: Affix ordering revised. *Language, 78* (3), 527-555.

Hayes, B., & Ma, T. (2005). Query Google. [Computer software]. Applet accessed July 14, 2005 at http://www.humnet.ucla.edu/humnet/linguistics/people/hayes/QueryGoogle/

Huang, Y-C. (2001). *Frequency effects on the processing of an aspect marker in Mandarin.* Unpublished master's thesis, National Chung Cheng University, Chiayi, Taiwan.

Hung, D. L., Tzeng, O. J. L. & Ho, C.-Y. (1999). Word superiority effect in the visual processing of Chinese. In O. J. L. Tzeng (Ed.) *Journal of Chinese Linguistics Monograph Series No. 13: The biological bases of language*, 61-95.

Katz, L., Rexer, K., & Lukatela, G. (1991). The processing of inflected words. *Psychological*

Research, 53, 25-32.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences.* New York: Brooks/Cole Publishing Company.

Labov, W. (1996). When intuitions fail. *Papers from the Regional Meetings, Chicago Linguistic Society, 32* (2), 77-105.

Li, C. N., & Thompson, S. A. (1976). The meaning and structure of complex sentences with *-zhe* in modern Mandarin. *Journal of the American Oriental Society, 96* (4), 512-519.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

Li, H., Li, T.-K., & Tseng, J.-F. (1997). *Guoyu cidian jianbianben bianji ziliao zicipin tongji baogao*. [Mandarin dictionary-based character and word frequency statistical report] Ministry of Education. Retrieved March 2, 2004, from http://140.111.1.22/clc/dict/htm/pin/start.htm

Li, P., Bates, E., & MacWhinney, B. (1993). Processing a language without inflections: A reaction time study of sentence interpretation in Chinese. *Journal of Memory and Language, 32,* 169-192.

Loftus, G. R., & Masson, M. E. J. (1994). Using confidence intervals in within-subject designs. *Psychonomic Bulletin & Review, 1*, 476-490.

Lorch, R. F., & Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16* (1), 149-157.

Ma, J. H. (1985). *A study of the Mandarin Chinese verb suffix zhe.* Taipei: Crane Publishing Company.

Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.) *Probabilistic linguistics* (pp. 289-341). Cambridge, MA: MIT Press.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology, 57* (3), 203-220.

Myers, J. (in press). Processing Chinese morphology: A survey of the literature. In G. Libben and G. Jarema (Eds.) *The representation and processing of compound words* (pp. 169-196). Oxford: Oxford University Press.

Myers, J., & Gong, S. (2002). Cross-morphemic predictability and the lexical access of compounds in Mandarin Chinese. *Folia Linguistica, 26* (1-2), 65-96.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht: Foris.

New, B., Brysbaert, M., Segui, J., Ferrand, L., & Rastle, K. (2004). The processing of singular and plural nouns in French and English. *Journal of Memory and Language, 51*, 568-585.

Packard, J. L. (2000). *The morphology of Chinese: A linguistic and cognitive approach.* Cambridge, UK: Cambridge University Press.

Pinker, S. (1999). *Words and rules: The ingredients of language.* New York: Basic Books.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416-426.

Ramscar, M. (2002). The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology, 45,* 45-94.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide*. Pittsburgh: Psychology Software Tools Inc.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37,* 118-139.

Sereno, J. A., & Jongman, A. (1997). Processing of English inflectional morphology. *Memory & Cognition, 2* (4), 425-437.

Shillcock, R. (1990). Lexical hypotheses in continuous speech. In G. T. M. Altmann (Ed.) *Cognitive models of speech processing: Psycholinguistic and computational perspectives* (pp. 24-49). Cambridge, MA: MIT Press.

Smith, C. S. (1991). *The parameter of aspect*. Dordrecht: Kluwer Academic Publishers.

Stump, G. T. (1998). Inflection. In A. Spencer and A. M. Zwicky (Eds.), *The handbook of morphology* (pp. 13-43). Oxford: Blackwell Publishers.

Taft, M. (1979). Recognition of affixed words and the word frequency effect. *Memory & Cognition, 7*, 263-272.

Taft, M. (2004). Morphological decomposition and the reverse base frequency effect. *The Quarterly Journal of Experimental Psychology, 57A* (4), 745-765.

Taft, M., Liu, Y., & Zhu, X. (1999). Morphemic processing in reading Chinese. In J. Wang, A. W. Inhoff, & H.-C. Chen (Eds.) *Reading Chinese script: A cognitive analysis* (pp. 91-113). Mahwah, NJ: Lawrence Erlbaum Associates.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge, UK: Cambridge University Press.

Ullman, M. T. (1999). Acceptability ratings of regular and irregular past-tense forms: Evidence for a dual-system model of language from word frequency and phonological neighbourhood effects. *Language and Cognitive Processes, 14* (1), 47-67.

Vogel Sosa, A., & MacFarlane, J. (2002). Evidence for frequency-based constituents in the mental lexicon: collocations involving the word *of*. *Brain and Language, 83*, 227-236.

Wheeldon, L. R., & Lahiri, A. (1997). Prosodic units in speech production. *Journal of Memory and Language*, *37*, 356-381.

Xue, N. (2001). *Defining and automatically identifying words in Chinese*. Unpublished doctoral thesis, University of Delaware, Newark, DE.

Zhou, X., & Marslen-Wilson, W. (1994). Words, morphemes, and syllables in the Chinese mental lexicon. *Language and Cognitive Processes, 9* (3), 393-422.

Zhou, X., & Marslen-Wilson, W. (2000). Lexical representation of compound words: Cross-linguistic evidence. *Psychologia, 43* (1), 47-66.

Zhou, X., Ostrin, R. K. & Tyler, L. K. (1993). The noun-verb problem and Chinese aphasia: comments on Bates et al. (1991). *Brain and Language, 45*, 86-93.

Zwicky, A. and Pullum, G. (1983). Cliticization vs. inflection: English "n't". *Language, 59,* 502-513.

TABLE 1

Analysis of variance results for Experiment 1

Group	Comparison		By participants		By items		Min *F′*	
			df	F_1	*df*	F_2	*df*	*minF′*
Bases	Words vs. nonwords	RT	1, 19	32.32*	1, 46	114.24*	1, 30	25.19*
		%	1, 19	7.92*	1, 46	4.44*	1, 64	2.84
	High vs. low frequency inflected forms	RT	1, 19	4.32	1, 22	1.26	1, 33	< 1
		%	1, 19	< 1	1, 22	< 1	1, 40	< 1
Inflected forms	Words vs. nonwords	RT	1, 19	14.46*	1, 46	71.44*	1, 27	12.03*
		%	1, 19	6.95*	1, 46	10.63*	1, 44	4.20*
	High vs. low frequency inflected forms	RT	1, 19	< 1	1, 22	< 1	1, 41	< 1
		%	1, 19	< 1	1, 22	< 1	1, 39	< 1

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 2

Summary of regression analysis for variables predicting reaction time in Experiment 1 (inflected forms)

Variable	*B*	*β*	*SE*	*t*
Log surface frequency	-4.12	-0.01	31.69	-0.13
Log base frequency	-143.08	-0.17	46.10	-3.10 *
Acceptability	-2.96	-0.00	31.06	-0.10

Note. *B* = mean by-participant raw regression coefficients, *β* = standardized regression coefficients, *SE* = standard errors for one-group *t* test conducted across by-participant regression coefficients.

* $p < .05$

TABLE 3

Analysis of variance results for combined results involving Experiments 1 and 3

Experiment(s)	Comparison		By participants		By items		Min *F'*	
			df	F_1	*df*	F_2	*df*	*minF'*
Experiment 1	Bases vs. inflected forms	RT	1, 38	< 1	1, 22	2.69	1, 43	< 1
		%	1, 38	< 1	1, 22	< 1	1, 34	< 1
	High vs. low frequency inflected forms	RT	1, 38	1.23	1, 22	< 1	1, 42	< 1
		%	1, 38	< 1	1, 22	< 1	1, 34	< 1
	Interaction	RT	1, 38	< 1	1, 22	1.12	1, 59	< 1
		%	1, 38	< 1	1, 22	< 1	1, 22	< 1
Experiments 1 and 3	Nonword-base foils vs. noun-base foils	RT	1, 38	< 1	1, 22	9.63*	1, 44	< 1
		%	1, 38	4.71*	1, 22	5.49*	1, 58	2.54
	High vs. low frequency inflected forms	RT	1, 38	2.06	1, 22	< 1	1, 32	< 1
		%	1, 38	< 1	1, 22	< 1	1, 55	< 1
	Interaction	RT	1, 38	3.12	1, 22	3.76	1, 58	1.70
		%	1, 38	< 1	1, 22	< 1	1, 49	< 1

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 4

Analysis of variance results for Experiment 2

Group	Comparison		By participants		By items		Min *F'*	
			df	F_1	*df*	F_2	*df*	*minF'*
Bases	Words vs. nonwords	RT	1, 19	21.55*	1, 46	65.70*	1, 32	16.23*
		%	1, 19	4.92*	1, 46	9.22*	1, 40	3.21
	High vs. low frequency base forms	RT	1, 19	71.47*	1, 22	13.22*	1, 30	11.16*
		%	1, 19	14.46*	1, 22	17.63*	1, 40	7.94*
Inflected forms	Words vs. nonwords	RT	1, 19	40.31*	1, 46	66.16*	1, 43	25.05*
		%	1, 19	8.88*	1, 46	8.20*	1, 56	4.26*
	High vs. low frequency base forms	RT	1, 19	20.17*	1, 22	9.53*	1, 38	6.47*
		%	1, 19	6.78*	1, 22	6.56*	1, 41	3.33

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 5

Summary of regression analysis for variables predicting reaction time in Experiment 2 (inflected forms)

Variable	*B*	*β*	*SE*	*t*
Log surface frequency	-32.16	-0.16	7.12	-4.52 *
Log base frequency	-77.04	-0.26	16.30	-4.72 *
Acceptability	-2.63	-0.02	5.72	-0.46

Note. *B* = mean by-participant raw regression coefficients, *β* = standardized regression coefficients, *SE* = standard errors for one-group *t* test conducted across by-participant regression coefficients.

* $p < .05$

TABLE 6

Analysis of variance results for combined results in Experiment 2

Comparison		By participants		By items		Min *F'*	
		df	F_1	*df*	F_2	*df*	*minF'*
Bases vs. inflected forms	RT	1, 38	2.26	1, 22	33.06*	1, 43	2.11
	%	1, 38	1. 17	1, 22	1.74	1, 60	< 1
High vs. low frequency inflected forms	RT	1, 38	60.14*	1, 22	12.07*	1, 31	10.05*
	%	1, 38	21.22*	1, 22	24.01*	1, 57	11.26*
Interaction	RT	1, 38	< 1	1, 22	< 1	1, 27	< 1
	%	1, 38	2.36	1, 22	2.61	1, 57	1.24

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 7

Analysis of variance results for Experiment 3

Comparison		By participants		By items		Min *F'*	
		df	F_1	*df*	F_2	*df*	*minF'*
Words vs. nonwords	RT	1, 19	26.45*	1, 46	11.85*	1, 65	8.18*
	%	1, 19	< 1	1, 46	< 1	1, 51	< 1
High vs. low frequency surface forms	RT	1, 19	6.10*	1, 22	4.01	1, 40	2.42
	%	1, 19	< 1	1, 46	< 1	1, 40	< 1

Note. RT: reaction time; %: accuracy. * *p* < .05

TABLE 8

Summary of regression analysis for variables predicting reaction time in Experiment 3

Variable	*B*	*β*	*SE*	*t*
Log surface frequency	-46.02	-0.18	15.93	-2.89 *
Log base frequency	-30.88	-0.06	21.33	-1.45
Acceptability	-48.95	-0.12	17.96	-2.73 *

Note. *B* = mean by-participant raw regression coefficients, *β* = standardized regression coefficients, *SE* = standard errors for one-group *t* test conducted across by-participant regression coefficients.

* *p* < .05

TABLE 9

Analysis of variance results for Experiment 4

Group	Comparison		By participants		By items		Min F'	
			df	F_1	*df*	F_2	*df*	*minF'*
Bases	Words vs. nonwords	RT	1, 19	30.80*	1, 46	34.91*	1, 51	16.36*
		%[a]						
	High vs. low acceptability inflected forms	RT	1, 19	< 1	1, 22	< 1	1, 34	< 1
		%	1, 19	< 1	1, 22	< 1	1, 27	< 1
Inflected forms	Words vs. nonwords	RT	1, 19	26.96*	1, 46	38.01*	1, 46	15.77*
		%	1, 19	< 1	1, 46	< 1	1, 53	< 1
	High vs. low acceptability inflected forms	RT	1, 19	< 1	1, 22	< 1	1, 25	< 1
		%	1, 19	1.20	1, 22	< 1	1, 36	< 1

Note. RT: reaction time; %: accuracy. * $p < .05$

[a] Accuracy rates are identical, so F cannot be calculated.

TABLE 10

Summary of regression analysis for variables predicting reaction time in Experiment 4 (inflected forms)

Variable	B	β	SE	t
Log surface frequency	-44.07	-0.11	13.16	-3.35 *
Log base frequency	-123.29	-0.21	24.82	-4.97 *
Acceptability	7.02	0.02	10.98	0.64

Note. B = mean by-participant raw regression coefficients, β= standardized regression coefficients, SE = standard errors for one-group t test conducted across by-participant regression coefficients.

* $p < .05$

TABLE 11

Analysis of variance results for combined results involving Experiments 4 and 5

Experiment(s)	Comparison		By participants		By items		Min *F′*	
			df	F_1	*df*	F_2	*df*	*minF′*
Experiment 4	Bases vs. inflected forms	RT	1, 38	< 1	1, 22	2.28	1, 48	< 1
		%	1, 38	2.45	1, 22	2.05	1, 53	1.12
	High vs. low acceptability forms	RT	1, 38	< 1	1, 22	< 1	1, 48	< 1
		%	1, 38	1.31	1, 22	< 1	1, 28	< 1
	Interaction	RT	1, 38	< 1	1, 22	< 1	1, 42	< 1
		%	1, 38	< 1	1, 22	< 1	1, 41	< 1
Experiments 4 and 5	Nonword-base foils vs. noun-base foils	RT	1, 38	2.95	1, 22	17.58*	1, 49	2.53
		%	1, 38	< 1	1, 22	1.69	1, 60	< 1
	High vs. low acceptability forms	RT	1, 38	1.77	1, 22	< 1	1, 39	< 1
		%	1, 38	2.20	1, 22	1.12	1, 51	< 1
	Interaction	RT	1, 38	3.47	1, 22	1.07	1, 36	< 1
		%	1, 38	7.38*	1, 22	5.43*	1, 50	3.13

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 12

Analysis of variance results for Experiment 5

Comparison		By participants		By items		Min F'	
		df	F_1	*df*	F_2	*df*	*minF'*
Words vs. nonwords	RT	1, 19	6.42*	1, 46	14.92*	1, 36	4.49*
	%	1, 19	< 1	1, 46	< 1	1, 47	< 1
High vs. low acceptability inflected forms	RT	1, 19	3.80	1, 22	3.69	1, 41	1.87
	%	1, 19	6.45*	1, 22	8.86*	1, 39	3.73

Note. RT: reaction time; %: accuracy. * $p < .05$

TABLE 13

Summary of regression analysis for variables predicting reaction time in Experiment 5

Variable	*B*	β	*SE*	*t*
Log surface frequency	-22.43	-0.05	15.78	-1.42
Log base frequency	-13.78	-0.02	28.12	-0.49
Acceptability	19.54	0.05	13.38	1.46

Note. *B* = mean by-participant raw regression coefficients, β = standardized regression coefficients, *SE* = standard errors for one-group *t* test conducted across by-participant regression coefficients. No factor had a significant effect.

Appendix

Materials for Experiment 1
Bases for items with high surface frequency *zhe* (著) forms: 包圍 bāowéi "encircle," 打量 dǎliáng "take measure of," 交織 jiāozhī "interweave," 伴隨 bànsúi "accompany, " 夾雜 jiázá "mingle," 抱持 bàochí "hold a belief," 閃動 shǎndòng "flash," 搖晃 yáohuàng "falter," 跟隨 gēnsuí "follow," 摸索 mōsuǒ "grope," 環繞 huánrào "surround," 蘊藏 yùncáng "store up."

Bases for items with low surface frequency *zhe* (著) forms: 交談 jiāotán "converse," 依靠 yīkào "rely on," 玩弄 wánnòng "play with," 珍藏 zhēncáng "collect valuables," 追隨 zhuīsuí "follow, accompany," 塗抹 túmǒ "smear," 感嘆 gǎntàn "sigh," 遍布 biànbù "spread all over," 漂浮 piāofú "float," 滿載 mǎnzài "load," 編織 biānzhī "weave," 襯托 chèntuō "serve as a foil to."

Materials for Experiment 2
High frequency base forms: 批判 pīpàn "criticize," 依靠 yīkào "rely on," 侵害 qīnhài "encroach on," 珍藏 zhēncáng "collect valuables," 記載 jìzǎi "record," 提倡 tíchàng "promote," 期盼 qípàn "look forward to," 發生 fāshēng "happen," 搭乘 dāchéng "embark," 運送 yùnsòng "transport," 擁抱 yǒngbào "embrace," 隱藏 yǐncáng "hide."

Low frequency base forms: 閃動 shǎndòng "flash," 打量 dǎliáng "take measure of," 摻雜 chānzá "mix together," 交融 jiāoróng "blend," 遍布 biànbù "spread all over," 搖晃 yáohuàng "falter," 扼殺 èshā "smother," 漂浮 piāofú "float," 夾雜 jiázá "mingle," 玩弄 wánnòng "play with," 滿載 mǎnzài "load," 襯托 chèntuo "serve as a foil to."

Materials for Experiments 4 and 5
Bases for items with high acceptability *zhe* (著) forms: 摻雜 chānzá "mix together," 交談 jiāotán "converse," 依靠 yīkào "rely on," 爭論 zhēnglùn "debate," 玩弄 wánnòng "play with," 追逐 zhuīzhú "chase," 追隨 zhuīsuí "follow, accompany," 期盼 qípàn "look forward to," 塗抹 túmǒ "smear," 漂浮 piāofú "float," 滿載 mǎnzài "load," 銜接 xiánjie "join together."

Bases for items with low acceptability *zhe* (著) forms: 抗拒 kàngjù "resist," 呼叫 hūjiào "shout," 珍藏 zhēncáng "collect valuables," 閃動 shǎndòng "flash," 搭乘 dāchéng "embark," 搖晃 yáohuàng "falter," 遍布 biànbù "spread all over," 摸索 mōsuǒ "grope," 編織 biānzhi "weave," 凝聚 níngjù "condense," 襯托 chèntuō "serve as a foil to," 讚美 zànměi "praise."

Foil items
Nonword bases for foils in Experiment 1: 公按 gōngàn, 助合 zhùhé, 卷趨 juǎnqu, 奉拒 fèngjù, 承始 chéngshǐ, 抽親 chōuqīn, 花願 huāyuàn, 修漫 xiūmàn, 除討 chútǎo, 帶轉 dàizhuǎn, 產奪 chǎnduó, 規做 guīzuò, 報製 bàozhì, 尊勞 zūnláo, 握聞 wòwén, 替修 tìxiū, 發演 fāyǎn, 對低 duìdī, 瘋賴 fēnglài, 廚續 chúxù, 檢創 jiǎnchuàng, 避陷 bìxiàn, 翻添 fāntiān, 警派 jǐngpài.

Nonword bases for foils in Experiments 2 and 4: 公按 gōngàn, 助合 zhùhé, 卷趨 juǎnqu 奉拒 fèngjù, 承始 chéngshǐ, 抽親 chōuqīn, 花願 huāyuàn, 挑長 tiāocháng, 重梭 chóngsuō, 修漫 xiūmàn, 除討 chútǎo, 產奪 chǎnduó, 規做 guīzuò, 閉陷 bìxiàn, 報製 bàozhì, 尊勞 zūnláo, 握聞 wòwén,替修 tìxiū, 發演 fāyǎn, 對低 duìdī, 瘋賴 fēnglài, 廚續 chúxù, 翻添 fāntiān, 警派 jǐngpài.

Bases for foils in Experiments 3 and 5: 港灣 gǎngwān "bay," 色筆 sèbǐ "color pen," 音頻 yīnpín "sound frequency," 軍營 jūnyíng "military camp," 星河 xīnghé "galaxy," 馬蹄 mǎtí "horse hoof," 病容 bìngróng "sickly look," 過時 gùoshí "old-fashioned," 貝殼 bèiké "shell," 睡眠 shùimián "sleeping," 糕餅 gāobǐng "round flat cake," 階級 jiējí "social class," 本能 běnnéng "instinct," 乳液 rǔyiè "lotion," 物體 wùtǐ "object," 材質 cáizhí "material," 吸盤 xīpán "sucking disc," 絕技 júejì "stunt," 範圍 fànwéi "range," 野花 yěhuā "wild flower," 船貨 chuánhuò "cargo," 路費 lùfèi "traveling expenses," 堡壘 bǎolěi "fort," 櫥窗 chúchuāng "display window."