1 Supplementary materials

In the interest of providing sufficient information to replicate our study, luminance, pre-processing steps, and analyses are discussed in more detail below.

1.1 Luminance

We controlled for luminance in various ways. First, all depicted objects were visually adjusted to be of equal size and placed in front of a white 300 x 300 pixel background. This image filled 7% of the 1280 x 1024 pixel screen. Luminance measurements in Adobe Photoshop CS6 for each pixel of each image (values range from 0 to 255) showed that the images had a mean luminance of 232 (range: 200 – 249, SD: 13.4). More than half of the pixels were white (i.e., the background color) in all images such that the median luminance was 255 for each one of them. Individual luminance values for each stimulus are provided in Table 3. Second, the rest of the screen was uniformly set to gray (RGB value: 179, 179, 179) for all stimuli, which resulted in the majority of the screen (93%) having an identical luminance value throughout the trials. Third, the eye tracking calibration period (30 seconds) provided ample time for the participants’ eyes to adjust to the ambient light. Furthermore, an adaptation period before the measurement was part of each trial. Prior to the critical word presentation images were shown in silence for 1000 ms. The last 100 ms were used as baseline for adjusting the following data points. Pupil dilation latencies to light are reported to vary between 150 and 400 ms for control participants (c.f., p. 435 in Holmqvist et al., 2011). Considering the luminance of the screen (image + background), the difference between the brightest and darkest image as measured by Photoshop lies at 3.36, which amounts to about 1.3% of the range of possible values (i.e., completely black to completely white). Ambient luminance in the testing room was not constant across participants although it was within each participant. Natural light was blocked during the test and a fluorescent
lamp provided light which was dimmed to a comfortable level for the participant. Apart from controlling the visual stimuli in the way outlined above, no other corrections were performed.

1.2 Data pre-processing

We transformed the Tobii output (T1750, ClearView) files to matrices to be analyzed by R (version 3.1.0, R Core Team, 2014). The pupil data consisted of the estimated absolute mean diameter in mm for each data point (approximately every 20 ms) over the period of a trial. Sudden brief changes in pupil diameter (more than 0.05 mm in 20 ms) that are considered to be artefacts produced by the eye tracker were excluded from further analyses. Missing points were linearly interpolated if the interval missing was not more than 400 ms (the maximum duration of typical blinks, Beatty & Lucero-Wagoner, 2000). Afterwards, left and right pupil size values were averaged (following Fritzsche & Höhle, 2015). The overall correlation between left and right pupil size was high for all participants ($M = .95, SD = 0.03$).

Since variations of the pictures' luminance values, the ambient light, and individual differences affect pupil size (Beatty & Lucero-Wagoner, 2000), mean pupil dilation was calculated on a trial-wise basis, i.e., each trial served as its own baseline. This was possible because in the first second of each trial, when the picture was presented in silence, participants’ eyes adjusted for that particular luminance (Beatty & Lucero-Wagoner, 2000). Specifically, we corrected for inter-subject and inter-trial variation by subtracting a silent baseline value (i.e., a mean value of a 100 ms interval before the onset of the auditory label). For this reason, we did not collect individual stimulus and ambient luminance values. Trials with no data points in the baseline interval were excluded from further analyses (1.7% of trials). Manipulating the duration of the baseline interval (20 ms and 500 ms) did not significantly affect the results.
1.3 Analyses

The linear mixed effects models were built so that their random structure was maximally specified (Barr, Levy, Scheepers, & Tily, 2013; Jaeger, Graff, Croft, & Pontillo, 2011). Each intercept and slope fitted by the model was adjusted by the effect of condition and neighborhood density nested in participants and by vocabulary size nested in items. Due to the possibility of overfitting and hence producing convergence errors, the model could only be computed when vocabulary size in the random structure was dichotomized.

Since the Helmert-coded levels of *featural distance* were collinear (as they should be, being nested within each other), the correlation term in the random effect structure in *featural distance* was removed (Jaeger et al., 2011). The most parsimonious model was chosen through comparisons using Likelihood Ratio Tests (Pinheiro & Bates, 2000) via the `anova` function from the `stats` package (R Core Team, 2014).

Apart from analyzing the mean pupil size change reported in the main text, we also assessed peak dilation. The raw data points were fitted with the `smooth.spline` function from the `stats` R package (R Core Team, 2014). The smoothing function was specified as follows: smoothing parameter – [-1.5, 1.5], absolute precision – 0.0004, relative precision – .08, maximal number of iterations – 500, smoothing method = `gam`. After inspecting the evolution of the pupillary curve in individual trials, we operationalized finding the peak dilation to be used in the analysis as follows. The first local maximum that exceeded 80% of the absolute maximum dilation was chosen as the peak dilation point. If no dilation point met this criterion among the first three local maxima, the largest one of the three maxima was chosen. Analyses parallel to those performed with mean dilation determined that mean and peak dilation measures were in agreement by showing the same tendencies ($\beta_1 = 0.036, \ SE = 0.021, t = 1.73, \beta_2 = 0.041, \ SE = 0.022, t = 1.87, \beta_3 = 0.021, \ SE = 0.029, t = 0.72$).

Time-course analyses (post-hoc cluster-based permutation tests: Maris & Oostenveld, 2007) explored when significant differences emerged between each condition pair.
(using the eyetrackingR package: Dink & Ferguson, 2016). First, individual paired sample $t$-tests found the significant ($p < .05$) $t$-values across the whole time frame. Second, clusters (e.g., contiguous significant $t$-values) were identified, for which a cluster-level $t$-value was calculated as the sum of all single sample $t$-values within the cluster. Third, the significance of cluster-level $t$-values were assessed by generating Monte Carlo distributions ($N = 2000$) thereof and determining the probability of their occurrence given the distribution. Those clusters whose $t$ statistic exceeded the threshold ($t = 2.64$, Bonferroni-corrected for multiple comparisons) were then tabulated for each contrast. With this method, the following significant contrasts were identified (using the time_cluster_data function): correct vs. one-feature change: 1700–2300 ms ($\sum t = -10.78$); correct vs. two-feature change: 1200–1300 ms ($\sum t = -1.53$), 1500–2900 ms ($\sum t = -35.94$); correct vs. three-feature change: 400–700 ms ($\sum t = -1.53$), 1700–2400 ms ($\sum t = -12.48$); one-feature change vs. two-feature change: 2100–2900 ms ($\sum t = -16.21$), one-feature change vs. three-feature change: 300–500 ms ($\sum t = -3.07$).

No significant clusters were identified in the two-feature change vs. three-feature change contrast. Comparable results (i.e., significant contrasts across all condition pairs except the two-feature change vs. three-feature change) were obtained when the function time_cluster_data was supplied with a formula containing a linear mixed effects model. Such results – significant differences between all levels of featural distance except the two-feature change vs. three-feature change – are consistent with the ones obtained by linear mixed effects models (described in the Results section). These results support our conclusions presented in the Discussion section.

To determine whether filler items were treated similarly to the correctly pronounced experimental items in the study, a separate analysis was carried out on a restricted data set containing only those two levels of condition. The condition (correct vs. filler) variable was then sum-coded and used as a fixed effect in a linear mixed effects model, which was otherwise identical to the ones run in previous analyses. As expected, no
reliable difference was detected between pupillary responses given to filler items and those of correct experimental items indicated by the non-significant likelihood ratio test statistic comparing the null model and the model containing the fixed effect ($\chi^2 = 0.028$, $p = .87$).

There is a long-standing debate surrounding the estimation (and efficacy) of effect sizes and p-values in linear mixed effects models, due to the fact that depending on random structure specification (and whether one adopts a Bayesian or a pseudo-Bayesian approach), explained variance and significance values may change dramatically (Nakagawa & Schielzeth, 2013). Regarding marginal effect size changes in the comparison of nested models, the method suggested by Nakagawa and Schielzeth (2013) yields the following results: 0.03% for introducing neighborhood density, 0.37% for vocabulary size and 1% for featural distance. Significance values are reported in the main text when comparing nested models with likelihood ratio tests (Pinheiro & Bates, 2000).
Table 2: List of fillers (rows 1–20) and items related to another study (rows 21–40, half of them correctly pronounced, half of them epenthesized in each version).

| Word (English) |
|----------------|
| Adler (eagle)  |
| Birne (pear)   |
| Ente (duck)    |
| Finger (ibid.) |
| Fuchs (fox)    |
| Hemd (shirt)   |
| Herz (heart)   |
| Hund (dog)     |
| Korb (basket)  |
| Lampe (lamp)   |
| Mantel (coat)  |
| Mond (moon)    |
| Mund (mouth)   |
| Pilz (mushroom)|
| Pinsel (brush) |
| Schachtel (box)|
| Torte (cake)   |
| Weste (vest)   |
| Wolke (cloud)  |
| Zebra (ibid.)  |
| G(ə)rass (grass)|
| g(ə)rau (gray)|
| g(ə)rün (green)|
| K(ə)rabe (crab)|
| K(ə)ran (crane)|
| K(ə)röte (toad)|
| Sch(ə)necke (snail)|
| S(ə)tein (stone)|
| S(ə)tock (stick)|
| S(ə)tuhl (chair)|
| B(ə)lau (blue)|
| B(ə)lume (flower)|
| B(ə)rot (bread)|
| C(ə)lown (ibd.)|
| F(ə)lasche (bottle)|
| F(ə)liege (fly)|
| F(ə)rosch (frog)|
| G(ə)las (glass)|
| K(ə)nief (knee)|
| Sch(ə)wein (pig)|
Table 3: Luminance values of experimental items

| Picture | Mean  | SD    | Median |
|---------|-------|-------|--------|
| Schaf   | 244.72| 34.16 | 255    |
| Sofa    | 199.97| 79.84 | 255    |
| Sonne   | 231.64| 37.02 | 255    |
| Suppe   | 244.3 | 29.68 | 255    |
| Tisch   | 228.38| 57.20 | 255    |
| Boot    | 226.22| 58.66 | 255    |
| Teddy   | 215.59| 73.15 | 255    |
| Dusche  | 232.36| 53.51 | 255    |
| Decke   | 231.93| 37.03 | 255    |
| Fuß     | 246.53| 25.00 | 255    |
| Fisch   | 241.42| 39.33 | 255    |
| Fahne   | 239.11| 50.06 | 255    |
| Bett    | 225.55| 59.40 | 255    |
| Pony    | 214.86| 74.98 | 255    |
| Kaffee  | 231.85| 54.76 | 255    |
| Buch    | 209.28| 78.58 | 255    |
| Baby    | 234.5 | 53.17 | 255    |
| Schere  | 248.41| 29.14 | 255    |
| Kamm    | 245.47| 38.76 | 255    |
| Käse    | 248.86| 16.43 | 255    |
References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. doi: 10.1016/j.jml.2012.11.001

Dink, J., & Ferguson, B. (2016). eyetrackingR [Computer software manual]. Retrieved from http://www.eyetracking-R.com (R package version 0.1.6)

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye-tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press. doi: 10.1080/17470218.2015.1098709

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology, 15*(2), 281–320. doi: 10.1515/lity.2011.021

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution, 4*(2), 133–142. doi: 10.1111/j.2041-210x.2012.00261.x

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in s and s-PLUS*. Springer New York. doi: 10.1007/978-1-4419-0318-1

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/