A. A Brief Introduction to Mixed-Effects Models

In mixed-effects models, additional variance components are introduced into the fixed-effects (traditional regression) structure. Broadly speaking, these can be thought of as associating different types of variance to different groupings – for example, by items or by subjects, with the remaining unexplained variance included as the error or "residual" term. These variance groupings can be applied to any term in the model: the intercept (e.g. allowing for the intercept to vary between test subjects, thus allowing for each subject to have a different baseline response) or to the various slopes (e.g. the differential response to experimental manipulations, thus accommodating different responses between subjects).

A.1. Selection of Random-Effects Structure

A major topic of debate in the application of mixed models to psycho- and neurolinguistic data is the structure of the random effects.

While Baayen et al. (2008) recommend forward selection of the random-effects structure, starting from the minimal intercepts-only structure, Barr et al. (2013) recommend backwards selection from the maximal random-effects structure, and Barr (2013) takes this suggestion one step further and suggests including *all* interactions in the random-effects structure. In practice, Barr et al. (2013)'s suggestion is somewhat problematic as complex random-effect structures are costly to compute and often fail to converge on real data sets (Bates et al., 2015a). Moreover, the backward selection procedure suggested by Barr et al. potentially leads to issues with overparameterization (Bates et al., 2015a). Another suggestion common to the mixed model literature is to follow the random-effects structure that best models the experimental design (see for example the GLMM wiki) and use a parsimonious model (Bates et al., 2015a).

A.2. Content of Model Summaries

The full model summary, as produced by `lme4` (Bates et al., 2015b) and formatted by `lmerOut` for LaTeX, includes the following components:

- Statement of fitting method – *maximum likelihood* or *REML* (*restricted* or *residual* maximum likelihood).
- Measures of fit, dependent on fitting method:
  - ML:
    - AIC: Akaike Information Criterion (Akaike, 1974) – goodness-of-fit measure penalized for number of parameters. Smaller is better, but there is no absolute "good" value.
    - BIC: Bayesian Information Criterion (Schwarz, 1978) – goodness-of-fit measure penalized for number of parameters. Similar to AIC, but with

larger penalty for number of parameters. Smaller is better, but there is no absolute “good” value.

    - logLik: Log-Likelihood – unpenalized goodness-of-fit measure. Always negative; larger (closer to zero) is better.
    - deviance – unpenalized goodness-of-fit measure, equal to -2 logLik. Similar to Residual Sum of Squares (RSS) in traditional linear regression. Smaller is better.

  - REML: REML criterion at convergence – similar to deviance but dependent on the fixed-effects parameterization

- Summary of variance components:
  - Scaled residuals – summary of the distribution of the final residual error in the model
  - Random effects – summary of the the different variance components by group on both the Variance and Standard Deviation scales. As residuals are technically a random effect, they are also listed here again.
  - Number of observations and the number of groups in each grouping term. For the subject-level grouping used here, the latter should match the number of subjects. The number of observations is equal to the number of subjects times the number of electrodes times the number of trials. In the case of the manipulations presented here with all content words, we have 52 subjects × 3 electrodes × 1682 epochs extracted = 262,392 observations.
- Summary of fixed effects – comparable to traditional regression output with:
  - Parameter estimates
  - Standard Error
  - t-values (ratio of estimate to error)
  - but no p-values, as it is non trivial to estimate the degrees of freedom in general. See the R FAQ 7.35 “Why are p-values not displayed when using lmer()?” for more information. The package `lmerTest` implements a variant where the degrees of freedom are estimated via the Satterthwaite approximation and the variances optionally corrected by the Kenward-Roger approximation. Beyond doubts about the accuracy of such approximations, these corrections are not necessary models for typical ERP datasets, as the effective degrees of freedom is always high enough that we can use the fact that the t distribution converges asymptotically to the normal distribution with increasing degrees of freedom.

A.3. Fitting Method

Several methods for fitting mixed-effects models exist based on different approximation methods. For the non-generalized linear case (linear mixed model, LMM), the three most common methods are residualized maximum likelihood (REML), maximum likelihood estimation (ML or MLE) and Markov-Chain Monte Carlo (MCMC) methods. The MCMC class of methods has been widely adopted in Bayesian settings and is extremely flexible and accurate in the limit, but is comparatively slow and complex and requires the specification of a prior (Bolker et al., 2009). Likelihood-based methods are faster and directly interpretable in a frequentist framework. In some sense, likelihood-based methods attempt to find the best model in terms of parameter values for a given specification and dataset: the general form of the model is specified by the experimenter and this is "fitted" to the data such that the probability that the data arises from the model ($P(D|M)$, i.e. the likelihood of the model given the data) is maximized.

Very coarsely, REML attempts to provide less biased estimates of the variance components (random-effects) in much the same way that Bessel's correction (using $n-1$ in the denominator instead of n) works for estimates of the population variance, $\hat{\sigma}^2$. One of the paradoxes of basic statistics is that Bessel's correction provides an unbiased estimate of the variance but a biased estimate of the standard deviation, $\hat{\sigma}$. This is related to the basic observation that ML estimators are invariant under many common transformations (e.g. logarithms und square roots) – the square root of the ML estimate for the variance is the MLE of the standard deviation, but unbiased non-ML estimators such as Bessel's correction, are not. Combined with the observation that the distribution of variance estimates is highly skewed and thus poorly summarized by traditional measures of location, and the necessity of unbiased estimates becomes questionable.[1]

Moreover, the unbiased estimator comes at a cost. The numerical formulation for REML leads to its "likelihood" being dependent on the (formulation of the) fixed-effects model matrix and as such REML-fitted models are not in general comparable via likelihood-ratio tests (see below). This again highlights a fundamental issue with REML: it is not actually *the* "likelihood", which is completely determined by the probability model (as noted by Douglas Bates on the R-SIG-ME mailing list), and as such it is somewhat misleading to label other optimization criteria as being some "form" of maximum likelihood.

Due to these issues with REML, full-model summaries and estimates of coefficients (including graphical presentations) are given using ML-fitted models. Nonetheless, REML-estimates are needed for computing F-tests with the Kenward-Roger variance correction and Satterthwaite degrees-of-freedom correction presented in the main text. The more anticonservative and much faster χ^2 tests can be computed with ML-estimates, in part because they are not dependent on variance estimates for the denominator.

A.4. Comparing Model Fit

It is often useful to determine which of several models best describes or "fits" the data. (Log) Likelihood provides a direct measure of model fit – likelihood is simply the probability that a given model would generate the exact dataset in question. (This is occasionally called the *marginal likelihood* of the model as it measures the likelihood of

[1] This is a brief summary of a longer comment by Douglas Bates on the R-SIG-ME mailing list

the model as a whole, i.e. over all conditions and terms, and not that of any particular effect.[2]) However, it is always possible to increase the likelihood by adding additional parameters, e.g. by adding a per-observation parameter, thus perfectly modelling the observed data. Additional parameters have two potential costs: (1) loss of parsimony and (2) loss of predictive power. Following Occam's Razor, we should find the most parsimonious solution; we also want to use our models to make inferences about not just the observed data but also about future data, and thus we want a model that maximizes predictive power (even at the cost of descriptive power, i.e. poorer fit on the observed data). Many measures have been suggested as a combined measure of model fit and model parsimony; two of the most popular for likelihood-based methods are the Akaike Information Criterion (AIC, Akaike, 1974) and the Bayesian Information Criterion (BIC, Schwarz, 1978). Both are based on the deviance (i.e. -2 logLik) penalized for the number of parameters, but BIC penalizes more heavily – the penalty for AIC increases linearly with number of parameters[3] while the penalty for BIC increases multiplicatively with the number of parameters and the (logarithm of the) number of observations[4]. There is no absolute nor bad for log likelihood, AIC or BIC, but in all cases the general rule of thumb is that closer to zero is better.

For nested models, there is also a significance test for comparing fits, as measured by likelihood, called the likelihood-ratio test (LRT). Likelihood ratios follow a χ^2 distribution asymptotically and can thus be compared using this distribution as the reference distribution. The χ^2 degrees of freedom is equal to the difference in model degrees of freedom. This is why this test is only valid for nested models – the difference in model degrees of freedom between non-nested models does not measure the restriction of one model to another (i.e. the creation of a simpler, special case), which is required for the asymptotic χ^2 distribution. Because the deviance is equal to twice the log likelihood, the χ^2 statistic for the LRT reduces to the difference between the deviances.[5] The LRT can be overly conservative when the value of a parameter is on the edge of its parameter space, e.g. when testing (the addition of) variance components whose estimates are near zero as variance is per definition non negative (Stram and Lee, 1994; Bates, 2010; Pinheiro and Bates, 2000).

A.5. Analogues to the Coefficient of Determination (R^2)

In simple linear regression, the coefficient of determination, R^2, can be interpreted as represented the percent of variance explained by the model. Conveniently, R^2 is the square of Pearson's correlation coefficient, r, for the correlation between the dependent and independent measures. When extended to multiple regression, R^2 can still be a useful tool but becomes more complex, in much the same way that correlation becomes more complex in a multivariate setting. Nonetheless, an adjusted (for parsimony/model

[2] See below for an explanation of the term "marginal".

[3] AIC $= -2\log L + 2p$, where L is the likelihood and p is the number of parameters

[4] BIC $= -2\log L + p\log n$, where L is the likelihood, p is the number of parameters and n is the number of observations

[5] Recall that $\log\frac{a}{b} = \log a - logb$.

complexity) R^2 value is still often reported for multiple regression. Both R^2 and adjusted R^2 have a number of useful properties and a relatively straightforward interpretation.

In a mixed-effects context, however, there is no measure with all the properties of R^2. Intuitively, this can be thought of as related to the complexity of correlation in a within-subjects design. Calculating correlation across trials without reference to subjects loses information and suffers from violations of independence. Calculating correlation within subjects and then averaging ignores issues related to Simpson's Paradox. Both methods have advantages and disadvantages, but neither has all the properties and ease of interpretation of bivariate correlation. Similarly, several pseudo R^2 measures have been proposed for mixed-effects models (e.g. Edwards et al., 2008; Gelman and Pardoe, 2006; Heinzl et al., 2005; Orelien and Edwards, 2008; Xu, 2003; Snijders and Bosker, 1994; Hssjer, 2008), but they do not enjoy the ease of interpretation of R^2 in simple regression, especially in mixed-effects models with multiple, interacting predictors, as has been noted by Douglas Bates on R-SIG-mixed-models and elsewhere (see the GLMM FAQ for a discussion of the issues involved as well as links to previous, longer discussion). Finally, as Payne et al. (2015) noted in their sentence-level analyses, where the noise is already lower than in a naturalistic context, the inter-trial variability of the EEG is such that pseudo R^2 measures do not offer much insight, even where the effects are strong.

A.6. Analysis of Deviance and Wald Tests for Linear Hypotheses

For large models, examining individual coefficients can be tedious and difficult. One possibility is repeated testing of nested models via likelihood-ratio tests, removing each predictor and its higher-level interactions one at a time to determine the (marginal) contribution of each predictor to model fit. This is a tedious and extremely computationally intensive process for models with multiple predictors.

Alternatively, we can use Type-II (marginal) Wald tests in an Analysis of Deviance. (Technically, the t tests on the coefficients are also Wald tests.) These tests measure the impact on the model of removing a particular term from the model, e.g. the change in deviance, and are, under certain assumptions, asymptotically equivalent to likelihood-ratio tests. As they measure the impact of a particular term (e.g. morphology) and not that of a particular coefficient or contrast (e.g. "nominative $>$ mean"), they also provide a more compact way of examining the effect of a given manipulation across contrast levels. This is analogous to the F-tests in a traditional Analysis of Variance, instead of examining of the coefficients of the linear model that ANOVA is based upon.

The naive and computationally simple approach is to assume "infinite" denominator degrees of freedom for the F tests in an Analysis of Deviance. This is equivalent to the assumption that the coefficients follow a normal and not a t distribution. This yields χ^2 tests where the degrees of freedom are the number of coefficients in the model for a given term (i.e. the same as in the LRT). A more accurate but much more computationally complex approach uses the Kenward-Roger corrected estimates for the variance and the Satterthwaite approximation for the degrees of freedom. When there are sufficient numbers of both observations and levels of the grouping variables (for the random effects), the two methods yield similar results, but generally the χ^2 test statistic yields anticonversative estimates compared to the F test statistic.

Type-II Wald tests have a number of problems (cf. Fox, 2016, pages 724–725, 737–738, and discussions on R-SIG-mixed-models), but even assuming that their results yield an anti-conservative estimate, they still allow of a much simpler summary of effect structure (cf. Bolker et al., 2009). A simple way to compensate for anti-conservative estimate is adopt a stricter significance threshold.

As the Wald tests presented here (Type-II) are *marginal tests*, they test the effect of completely removing a given term – and thus all of its interactions – from the model. As such, it is possible that the Wald test for a lower-order effect (e.g. main effect) is significant, although the corresponding t-value in the summary fails to achieve the $|t| \geq 2$ threshold. Since it is problematic to interpret main effects in the presence of interactions anyway, this is not a large problem (cf. Venables, 1998).[6]

B. Full Model Summaries and Additional Type-II Wald Tests

In order to make the models and their fits more readily comparable with each other, all models were estimated with Maximum Likelihood estimation (ML, i.e. with `REML=FALSE` in `lme4`, cf. Pinheiro and Bates, 2000; Baayen et al., 2008; Bates et al., 2015b). For the Wald F-tests here and in the main text (computed with `car::Anova()`, Fox and Weisberg, 2011), it was necessary to refit these models with REML for the test in order to perform the necessary operations on the variance components.. This not problematic: REML and ML yield similar results in models with large numbers of observations (Fox, 2016). The original calculation with the less accurate Wald χ^2 test statistic yielded the same pattern of results for both ML and REML fits.

For the model summaries, we view $|t| > 2$ (i.e., the estimate of the coefficient is more than twice as large as the error in the estimate) as being indicative of a reliable estimate in the sense that the estimate is distinguishable from noise. We view $|t| < 2$ as being unreliable estimates, which may be an indicator of low power or of a generally trivial effect. (We note that Baayen et al. (2008) use $|t| > 2$ as approximating the 5%-significance level.)

For the Type-II Wald tests, we use the p-values as a rough indication (see above) of reliability of the estimate across contrasts, which each receive their own coefficient in the model, and a quick way to get an overview of overall model structure.

C. Software Version (R Session Information)

```
R version 3.3.0 (2016-05-03)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.10.5 (Yosemite)
```

[6] Indeed, this is how Type-II (marginal) – tests differ from Type-I (sequential) and Type-III (non-sequential, i.e. not sensitive to order and keeping higher-order interactions when testing a lower-order term).

Table S1: Summary of model fit for (corpus) frequency class and index (ordinal position) in the time window 300–500ms from stimulus onset using all content words. Neither the main effect for index nor interaction term yields a reliable estimate.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance	
2021954	2022017	-1010971	2021942	
Scaled residuals:				
Min	1Q	Median	3Q	Max
-33.06	-0.53	0	0.53	38.43
Random effects:				
Groups	Name	Variance	Std.Dev	
subj	(Intercept)	0.10	0.31	
Residual		130.01	11.40	
Number of obs: 262392, groups: subj, 52.				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	0.54	0.13	4.1	
index	−6e−05	0.00013	−0.46	
corpus.freq	−0.068	0.0089	−7.6	
index:corpus.freq	1e−05	9.5e−06	1.1	

```
locale:
[1] en_AU.UTF-8/en_AU.UTF-8/en_AU.UTF-8/C/en_AU.UTF-8/en_AU.UTF-8

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] lmerOut_0.1     reshape2_1.4.1  plyr_1.8.4      lme4_1.1-12
[5] Matrix_1.2-6    lattice_0.20-33 effects_3.1-1   car_2.1-2

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.5        MASS_7.3-45        grid_3.3.0         xtable_1.8-2
 [5] nlme_3.1-128       MatrixModels_0.4-1 magrittr_1.5       stringi_1.1.1
 [9] SparseM_1.7        minqa_1.2.4        nloptr_1.0.4       splines_3.3.0
[13] tools_3.3.0        stringr_1.0.0      pbkrtest_0.4-6     parallel_3.3.0
[17] colorspace_1.2-6   mgcv_1.8-12        nnet_7.3-12        quantreg_5.26
```

Table S2: Comparison of models for (corpus) frequency class with and without index (ordinal position). Including index does not significantly improve model fit as evidenced by both the likelihood-ratio test and the information criteria.

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr(¿χ^2)
m.freq	4	2021954	2021995	-1010973	2021946			
m.freq.index	6	2021954	2022017	-1010971	2021942	3.74	2	0.154

Table S3: Summary of model fit for relative frequency class in the time window 300–500ms from stimulus onset using all content words.

Linear mixed model fit by maximum likelihood				
AIC	BIC	logLik	deviance	
2022083	2022125	-1011038	2022075	
Scaled residuals:				
Min	1Q	Median	3Q	Max
-33.08	-0.53	0	0.53	38.4
Random effects:				
Groups	Name	Variance	Std.Dev	
subj	(Intercept)	0.10	0.31	
Residual		130.08	11.41	
Number of obs: 262392, groups: subj, 52.				
Fixed effects:				
	Estimate	Std. Error	t value	
(Intercept)	0.25	0.091	2.8	
rel.freq	-0.089	0.013	-6.6	

References

Akaike, H., dec 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Baayen, R. H., Davidson, D. J., Bates, D. M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language 59, 390–412.

Barr, D. J., 2013. Random effects structure for testing interactions in linear mixed-effects models. Frontiers in Psychology 4 (328).

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, 255–278.

Table S4: Comparison of models for relative frequency class with and without index (ordinal position). Including index significantly improves model fit as evidenced by both the likelihood-ratio test and AIC.

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr(¿χ^2)
m.rel	4	2022083	2022125	-1011037	2022075			
m.rel.index	6	2022078	2022141	-1011033	2022066	8.61	2	0.0135

Bates, D., Kliegl, R., Vasishth, S., Baayen, H., 2015a. Parsimonius mixed models. arXiv, 1506.04967v1.

Bates, D., Maechler, M., Bolker, B. M., Walker, S., 2015b. Fitting linear mixed-effects models using lme4. arXiv, 1406.5823.

Bates, D. M., 2010. lme4: Mixed-effects modeling with R. Draft.
URL `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., White, J.-S. S., Mar 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol 24 (3), 127–35.

Edwards, L. J., Muller, K. E., Wolfinger, R. D., Qaqish, B. F., Schabenberger, O., 2008. An r^2 statistic for fixed effects in the linear mixed model. Statistics in Medicine 27 (29), 6137–6157.

Fox, J., 2016. Applied Regression Analysis and Generalized Linear Models, 3rd Edition. Sage, Thousand Oaks, CA.

Fox, J., Weisberg, S., 2011. An R Companion to Applied Regression, 2nd Edition. Sage, Thousand Oaks CA.
URL `http://socserv.socsci.mcmaster.ca/jfox/Books/Companion`

Gelman, A., Pardoe, I., 2015/09/30 2006. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. Technometrics 48 (2), 241–251.

Heinzl, H., Waldhr, T., Mittlbck, M., 2005. Careful use of pseudo r-squared measures in epidemiological studies. Statistics in Medicine 24 (18), 2867–2872.

Hssjer, O., 10 2008. On the coefficient of determination for mixed regression models. Journal of Statistical Planning and Inference 138 (10), 3022–3038.

Orelien, J. G., Edwards, L. J., 1 2008. Fixed-effect variable selection in linear mixed models using r^2 statistics. Computational Statistics and Data Analysis 52 (4), 1896–1907.

Payne, B. R., Lee, C.-L., Federmeier, K. D., 2015. Revisiting the incremental effects of context on word processing: Evidence from single-word event-related brain potentials. Psychophysiology.
URL http://dx.doi.org/10.1111/psyp.12515

Pinheiro, J., Bates, D., 2000. Mixed-Effects Models in S and S-PLUS. Springer New York.
URL https://books.google.de/books?id=3TVDAAAAQBAJ

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6 (2), 461–464.

Snijders, T. A. B., Bosker, R. J., 1994. Modeled variance in two-level models. Sociological Methods and Research 22 (3), 342–363.

Stram, D. O., Lee, J. W., 12 1994. Variance components testing in the longitudinal mixed effects model. Biometrics 50 (4), 1171–1177.

Venables, W. N., October 1998. Exegeses on linear models. In: S-PLUS User's Conference. Washington, DC.

Xu, R., 2003. Measuring explained variation in linear mixed effects models. Statistics in Medicine 22 (22), 3527–3541.

Table S5: Summary of model fit for index and linguistic cues (animacy, morphology, linear position) known to elicit N400-like effects. Dependent variable are single-trial means in the time window 300–500ms from stimulus onset using only subjects and (direct) objects. For animacy and position, the coefficients are named for the dispreferred condition and represent the contrast “dispreferred ¿ preferred”. Morphology also has an additional ‘neutral’ level for ambiguous case marking, and so the coefficients represent the contrast to that level. Scaled deviation (sum) encoding was used so that the coefficients are directly interpretable as the difference between means in the given contrast.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
530393	530630	-265170	530341

Scaled residuals:

Min	1Q	Median	3Q	Max
-11.85	-0.54	0	0.54	12.97

Random effects:

Groups	Name	Variance	Std.Dev
subj	(Intercept)	0.20	0.44
Residual		125.88	11.22

Number of obs: 69108, groups: subj, 52.

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	−0.2	0.16	−1.3
index	−0.00037	0.00018	−2.1
inanimate	0.21	0.29	0.74
accusative	−0.26	0.4	−0.64
nominative	1.1	0.48	2.2
non-initial	−0.63	0.29	−2.2
index:inanimate	−0.00058	0.00036	−1.6
index:accusative	0.0014	0.00046	3.1
index:nominative	−0.0026	0.00064	−4.1
inanimate:accusative	−1.1	0.8	−1.3
inanimate:nominative	0.89	0.96	0.93
index:non-initial	−0.00024	0.00036	−0.66
inanimate:non-initial	0.86	0.57	1.5
accusative:non-initial	0.6	0.8	0.74
nominative:non-initial	−0.68	0.96	−0.71
index:inanimate:accusative	0.0022	0.00091	2.5
index:inanimate:nominative	−0.0022	0.0013	−1.7
index:inanimate:non-initial	−0.0014	0.00071	−1.9
index:accusative:non-initial	0.0022	0.00091	2.5
index:nominative:non-initial	−0.0039	0.0013	−3.1
inanimate:accusative:non-initial	−1.3	1.6	−0.81
inanimate:nominative:non-initial	4.9	1.9	2.6
index:inanimate:accusative:non-initial	0.00021	0.0018	0.12
index:inanimate:nominative:non-initial	−0.0054	0.0026	−2.1

Table S6: Type-II Wald tests for the clearest effects in the model combining index, (corpus) frequency and linguistic cues.

	F	Df	Df.res	Pr(¿F)
index	6.83	1	69009	0.00895
corpus.freq	55.03	1	69009	< 0.001
morphology	19.65	2	69009	< 0.001
position	3.15	2	69009	0.0429
corpus.freq:morphology	5.51	2	69009	0.00405
index:position	8.84	1	69009	0.00294
corpus.freq:position	19.90	1	69009	< 0.001
morphology:position	11.69	2	69009	< 0.001
index:animacy:morphology	3.81	2	69009	0.0222
index:corpus.freq:position	4.47	1	69009	0.0345
index:animacy:position	6.01	1	69009	0.0143
corpus.freq:morphology:position	5.89	2	69009	0.00277
animacy:morphology:position	3.46	2	69009	0.0314
index:corpus.freq:animacy:morphology	8.36	2	69009	< 0.001
corpus.freq:animacy:morphology:position	4.75	2	69009	0.00866

Table S7: Summary of model fit for index, (corpus) frequency and linguistic cues (animacy, morphology, linear position) known to elicit N400-like effects. Dependent variable are single-trial means in the time window 300–500ms from stimulus onset using only subjects and (direct) objects. For animacy and position, the coefficients are named for the dispreferred condition and represent the contrast “dispreferred ¿ preferred”. Morphology also has an additional ‘neutral’ level for ambiguous case marking, and so the coefficients represent the contrast to that level. Scaled deviation (sum) encoding was used so that the coefficients are directly interpretable as the difference between means in the given contrast.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
530282	530739	-265091	530182

Scaled residuals:

Min	1Q	Median	3Q	Max
-11.89	-0.54	0	0.54	12.99

Random effects:

Groups	Name	Variance	Std.Dev
subj	(Intercept)	0.20	0.44
Residual		125.59	11.21

Number of obs: 69108, groups: subj, 52.

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	0.88	0.54	1.6
index	0.00084	0.00073	1.2
corpus.freq	−0.089	0.042	−2.1
inanimate	−0.41	1.1	−0.38
accusative	0.41	1.6	0.27
nominative	0.039	1.8	0.022
non-initial	−2.8	1.1	−2.6
index:corpus.freq	−6.3e−05	5.9e−05	−1.1
index:inanimate	0.003	0.0015	2.1
corpus.freq:inanimate	0.069	0.084	0.82
index:accusative	0.00086	0.0021	0.42
index:nominative	0.0016	0.0025	0.62
corpus.freq:accusative	0.082	0.12	0.7
corpus.freq:nominative	−0.047	0.14	−0.33
inanimate:accusative	11	3.1	3.6
inanimate:nominative	−11	3.5	−3.2
index:non-initial	0.0024	0.0015	1.6
corpus.freq:non-initial	0.14	0.084	1.7
inanimate:non-initial	−1	2.2	−0.48
accusative:non-initial	−0.38	3.1	−0.12
nominative:non-initial	2.4	3.5	0.69
index:corpus.freq:inanimate	−0.00022	0.00012	−1.9
index:corpus.freq:accusative	−0.0001	0.00017	−0.59
index:corpus.freq:nominative	−0.00013	0.00019	−0.66
index:inanimate:accusative	−0.013	0.0041	−3.2
index:inanimate:nominative	0.019	0.005	3.7
corpus.freq:inanimate:accusative	−0.82	0.23	−3.5
corpus.freq:inanimate:nominative	0.83	0.28	2.9
index:corpus.freq:non-initial	−0.00016	0.00012	−1.3
index:inanimate:non-initial	2.8e−05	0.0029	0.0096
corpus.freq:inanimate:non-initial	0.022	0.17	0.13
index:accusative:non-initial	−0.00014	0.0041	−0.035
index:nominative:non-initial	−5.5e−05	0.005	−0.011
corpus.freq:accusative:non-initial	−0.084	0.23	−0.36
corpus.freq:nominative:non-initial	−0.17	0.28	−0.6
inanimate:accusative:non-initial	−14	6.2	−2.3
inanimate:nominative:non-initial	11	7.1	1.5
index:corpus.freq:inanimate:accusative	0.0011	0.00035	3.2
index:corpus.freq:inanimate:nominative	−0.0014	0.00038	−3.6
index:corpus.freq:inanimate:non-initial	4.1e−05	0.00023	0.18
index:corpus.freq:accusative:non-initial	0.00021	0.00035	0.59

Table S8: Model comparison for linguistic-cue based models with index and (corpus) frequency. Index and frequency significantly improve model fit as measured by the likelihood-ratio test and AIC.

	Df	AIC	BIC	logLik	deviance	χ^2	χ^2 Df	Pr(¿χ^2)
prom	14	530425	530553	-265198	530397			
prom.index	26	530392	530630	-265170	530340	56.43	12	< 0.001
prom.freq.index	50	530281	530738	-265090	530181	159.18	24	< 0.001

Table S9: Type-II Wald tests for the clearest effects in the model combining orthographic length, (corpus) frequency and linguistic cues.

	F	Df	Df.res	Pr(¿F)
corpus.freq	9.85	1	69009	0.0017
morphology	22.21	2	69009	< 0.001
position	8.39	1	69009	0.00378
ortho.len:morphology	3.02	2	69009	0.049
corpus.freq:morphology	5.35	2	69009	0.00476
corpus.freq:position	6.02	1	69009	0.0141
morphology:position	19.96	2	69009	< 0.001
ortho.len:corpus.freq:animacy	15.96	1	69009	< 0.001
ortho.len:corpus.freq:morphology	3.47	2	69009	0.0312
ortho.len:corpus.freq:position	10.38	1	69009	0.00127
ortho.len:animacy:position	6.52	1	69009	0.0107
ortho.len:corpus.freq:animacy:morphology	24.50	2	69009	< 0.001
ortho.len:corpus.freq:animacy:position	34.30	1	69009	< 0.001
corpus.freq:animacy:morphology:position	4.28	2	69009	0.0139
ortho.len:corpus.freq:animacy:morphology:position	4.23	2	69009	0.0145

Table S10: Summary of model combining linguistic cues with corpus frequency and orthographic length. Dependent variable are single-trial means in the time window 300–500ms from stimulus onset using only subjects and (direct) objects. For animacy and position, the coefficients are named for the dispreferred condition and represent the contrast "dispreferred ¿ preferred". Morphology also has an additional 'neutral' level for ambiguous case marking, and so the coefficients represent the contrast to that level. Scaled deviation (sum) encoding was used so that the coefficients are directly interpretable as the differenc e between means in the given contrast.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
530228	530686	-265064	530128

Scaled residuals:

Min	1Q	Median	3Q	Max
-11.78	-0.54	0	0.54	12.94

Random effects:

Groups	Name	Variance	Std.Dev
subj	(Intercept)	0.20	0.45
Residual		125.50	11.20

Number of obs: 69108, groups: subj, 52.

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	3.6	1.7	2.1
ortho.len	−0.4	0.28	−1.4
corpus.freq	−0.23	0.1	−2.3
inanimate	1.8	3.4	0.52
accusative	−5.7	6.2	−0.93
nominative	9.1	4.5	2
non-initial	3.8	3.4	1.1
ortho.len:corpus.freq	0.02	0.014	1.4
ortho.len:inanimate	−0.069	0.57	−0.12
corpus.freq:inanimate	−0.1	0.2	−0.5
ortho.len:accusative	1.1	1.1	1
ortho.len:nominative	−1.4	0.68	−2
corpus.freq:accusative	0.43	0.32	1.3
corpus.freq:nominative	−0.55	0.31	−1.8
inanimate:accusative	12	12	0.97
inanimate:nominative	4.3	8.9	0.48
ortho.len:non-initial	−0.64	0.57	−1.1
corpus.freq:non-initial	−0.29	0.2	−1.5
inanimate:non-initial	−4.2	6.8	−0.61
accusative:non-initial	−9	12	−0.73
nominative:non-initial	18	8.9	2
ortho.len:corpus.freq:inanimate	0.0039	0.028	0.14
ortho.len:corpus.freq:accusative	−0.063	0.048	−1.3
ortho.len:corpus.freq:nominative	0.073	0.038	1.9
ortho.len:inanimate:accusative	−2	2.1	−0.94
ortho.len:inanimate:nominative	0.43	1.4	0.32
corpus.freq:inanimate:accusative	−0.43	0.65	−0.66
corpus.freq:inanimate:nominative	−0.87	0.61	−1.4
ortho.len:corpus.freq:non-initial	0.039	0.028	1.4
ortho.len:inanimate:non-initial	0.88	1.1	0.78
corpus.freq:inanimate:non-initial	−0.018	0.4	−0.045
ortho.len:accusative:non-initial	1.4	2.1	0.65
ortho.len:nominative:non-initial	−2.5	1.4	−1.8
corpus.freq:accusative:non-initial	0.55	0.65	0.85
corpus.freq:nominative:non-initial	−1.1	0.61	−1.8
inanimate:accusative:non-initial	−14	25	−0.58
inanimate:nominative:non-initial	40	18	2.2
ortho.len:corpus.freq:inanimate:accusative	0.081	0.096	0.84
ortho.len:corpus.freq:inanimate:nominative	0.043	0.075	0.57
ortho.len:corpus.freq:inanimate:non-initial	−0.019	0.055	−0.35
ortho.len:corpus.freq:accusative:non-initial	−0.069	0.096	−0.72
ortho.len:corpus.freq:nominative:non-initial	0.12	0.075	1.6

Table S11: Type-II Wald tests for the clearest effects in the model combining linguistic cues with both corpus and relative frequency.

	F	Df	Df.res	Pr(¿F)
rel.freq	3.85	1	69011	0.0496
corpus.freq	40.62	1	69011	< 0.001
morphology	18.51	2	69011	< 0.001
rel.freq:corpus.freq	9.30	1	69011	0.00229
rel.freq:animacy	10.51	1	69011	0.00119
corpus.freq:morphology	9.76	2	69011	< 0.001
animacy:morphology	3.83	2	69011	0.0217
corpus.freq:position	10.12	1	69011	0.00147
morphology:position	15.57	2	69011	< 0.001
rel.freq:corpus.freq:animacy	13.26	1	69011	< 0.001
rel.freq:corpus.freq:morphology	6.74	2	69011	0.00119
rel.freq:animacy:morphology	12.40	2	69011	< 0.001
rel.freq:morphology:position	8.98	2	69011	< 0.001
corpus.freq:morphology:position	5.41	2	69011	0.00449
animacy:morphology:position	4.19	2	69011	0.0152
rel.freq:corpus.freq:animacy:morphology	6.81	2	69011	0.0011
rel.freq:corpus.freq:animacy:position	25.65	1	69011	< 0.001
corpus.freq:animacy:morphology:position	6.88	2	69011	0.00103

Table S12: Summary of model combining linguistic cues with corpus and relative frequency. Dependent variable are single-trial means in the time window 300–500ms from stimulus onset using only subjects and (direct) objects. For animacy and position, the coefficients are named for the dispreferred condition and represent the contrast "dispreferred ¿ preferred". Morphology also has an additional 'neutral' level for ambiguous case marking, and so the coefficients represent the contrast to that level. Scaled deviation (sum) encoding was used so that the coefficients are directly interpretable as the difference between means in the given contrast.

Linear mixed model fit by maximum likelihood

AIC	BIC	logLik	deviance
530242	530681	-265073	530146

Scaled residuals:

Min	1Q	Median	3Q	Max
-11.87	-0.54	0	0.54	13

Random effects:

Groups	Name	Variance	Std.Dev
subj	(Intercept)	0.20	0.45
Residual		125.53	11.20

Number of obs: 69108, groups: subj, 52.

Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	−1.5	4.3	−0.36
rel.freq	0.4	0.64	0.63
corpus.freq	0.19	0.35	0.55
inanimate	−4.3	6.4	−0.67
accusative	11	14	0.8
nominative	−22	13	−1.7
non-initial	6.5	7.2	0.9
rel.freq:corpus.freq	−0.044	0.051	−0.87
rel.freq:inanimate	0.84	0.95	0.89
corpus.freq:inanimate	0.37	0.52	0.71
rel.freq:accusative	−1.3	2.1	−0.64
rel.freq:nominative	3.5	2	1.8
corpus.freq:accusative	−0.89	1.1	−0.8
corpus.freq:nominative	2	1.1	1.8
inanimate:accusative	46	15	3
inanimate:nominative	−46	25	−1.9
rel.freq:non-initial	−1.2	1.1	−1.1
corpus.freq:non-initial	−0.51	0.59	−0.86
inanimate:non-initial	−24	5.5	−4.5
accusative:non-initial	−9.7	26	−0.37
nominative:non-initial	30	18	1.7
rel.freq:corpus.freq:inanimate	−0.065	0.076	−0.85
rel.freq:corpus.freq:accusative	0.12	0.16	0.77
rel.freq:corpus.freq:nominative	−0.32	0.16	−2
rel.freq:inanimate:accusative	−6.9	2.3	−2.9
rel.freq:inanimate:nominative	7.2	3.7	2
corpus.freq:inanimate:accusative	−3	1.2	−2.5
corpus.freq:inanimate:nominative	2.8	2.1	1.4
rel.freq:corpus.freq:non-initial	0.085	0.087	0.97
rel.freq:inanimate:non-initial	4.1	0.88	4.7
corpus.freq:inanimate:non-initial	1.7	0.39	4.4
rel.freq:accusative:non-initial	1.5	4	0.38
rel.freq:nominative:non-initial	−4.5	2.8	−1.6
corpus.freq:accusative:non-initial	0.37	2.1	0.18
corpus.freq:nominative:non-initial	−1.9	1.5	−1.3
inanimate:accusative:non-initial	−19	11	−1.7
inanimate:nominative:non-initial	12	8.4	1.5
rel.freq:corpus.freq:inanimate:accusative	0.45	0.17	2.6
rel.freq:corpus.freq:inanimate:nominative	−0.41	0.3	−1.4
rel.freq:corpus.freq:inanimate:non-initial	−0.28	0.056	−5.1
rel.freq:corpus.freq:accusative:non-initial	−0.055	0.31	−0.18
rel.freq:corpus.freq:nominative:non-initial	0.26	0.23	1.1