Beyond Rule versus Rote? Processing of Distinctive Dative and Genitive Case-Markers in German

Christian Bentz (cb696@cam.ac.uk)
Department of Theoretical and Applied Linguistics, Sidgwick Avenue
Cambridge, CB3 9DA UK

Abstract

The rule versus rote distinction is one of the most debated issues in recent psycholinguistics. Dual route accounts hold that words can either be stored whole in the mental lexicon or computationally derived by simple combinatorial rules such as *stem+affix*. Within this framework, response latencies in lexical decision tasks have been applied to point out the difference between rote memorization, on the one hand, and combinatorial rule manipulation, on the other. However, this paper argues that there may be alternatives to this distinction. It will be shown that German nouns, which can be distinctively marked for number, case or both number and case, do elicit differing reaction times. Crucially, this effect can neither be explained by surface frequency effects nor by internal morphological structure. Rather, it seems to be triggered by the degree of embedding into usage-based units.

Keywords: Rule versus rote; lexical decision; German case-marking; usage-based units.

Introduction

The *rule* versus *rote* distinction in psycholinguistic theories of lexical access has been fiercely debated (see Pinker & Ullman, 2002 as well as McClelland & Patterson, 2002 for a review). Lexical decision tasks (LDT), priming studies, event related potentials and fMRI studies (see Clahsen, 1999 for a review) have been applied to answer the question whether lexical processing of morphologically simplex and complex items is rule-governed or associative, or both. It has been argued that lexical decision latencies can help us to distinguish processes involving abstract rule manipulation from mere memorization effects (Pinker & Ullman, 2002; Taft, 2004; Clahsen, 1999, Clahsen et al., 1997; Marslen-Wilson & Tyler, 2007, Sonnenstuhl & Huth, 2002). In this context, the absence of frequency effects for regularly derived forms has been explained by abstract *rule manipulation*, whereas the occurrence of frequency effects was associated with *rote memorization* of irregular forms (see for example Clahsen 1999: 998, but also Hahn & Nakisa 2000 for critical points). If these assumptions hold, then processing difficulties in lexical decision tasks must stem from:

a) The low frequencies of test items (in the case of memorization);

b) The difficulty of parsing by means of grammatical rules applied to derive the internal structure of a morphologically complex word (symbol manipulation).

However, the study presented here suggests that the 'grammatical load' of inflections is another potential factor relevant for processing difficulty, depending on *word external* rather than *word internal* factors. Along those lines, it will be argued that a usage-based account of lexical access can provide an alternative explanation of the processing difficulties reflected in lexical decision tasks

To this end, a lexical decision experiments was designed which involved German words with *-(e)n* and *-s* plural marking, which can optionally also encode dative and genitive case. It will be shown that forms with more *grammatical load*, i.e. forms encoding both case and plural meaning, elicited significantly longer response latencies than unmarked forms. Crucially, these prolonged latencies can neither be explained by token frequency effects nor by word-internal parsing, rather, the participants seemed to have invoked redundantly case-marked articles or prepositional phrases triggering case-marking. This way they could decide whether the case-marked word is a possible word form in German. This strategy prolongs reaction times (RTs) for morphologically complex forms.

Therefore, this paper will argue that the distinction between rule governed processes and memorization effects in LDT research lacks an important aspect of language processing: the embedding of items in phrases and sentences, i.e. usage-based units. In the following, the case-marking and plural paradigms of nouns in German will be sketched in section 1. In section 2 the methods and results of the LDT will be presented and discussed in section 3.

1. German Dative and Genitive Inflections

It is generally assumed that German has four distinct case-marking paradigms: *nominative*, *accusative*, *genitive* and *dative* (Engel, 1991: 505; Griesbach, 1986: 294; Kempe & MacWhinney 1998: 549). However, from the perspective of the noun declension the picture is more complicated. Since there is a fair amount of syncretism between case-markers and singular/plural markers across different noun classes, the only markers that are *distinctive inflectional case-markers*[1] are the *-(e)s genitive marker* for a subset of *singular* masculine and neuter nouns as well as the *-(e)n dative marker* in the *plural* for all genders (Griesbach, 1986: 294; Engel, 1991: 505). Hence, distinctively case marked forms are restricted to these *-(e)s* and *-n* inflections for some nouns.

[1] Inflectional markers that are distinct from the other plural or singular forms of the same declension class and hence clearly identify the surface form as case-marked for a specific case.

For example, the high frequent noun *Haus* (house) is inflected as *Häus-er* (houses) for all plural forms except for the dative, for which *in den Häus-er-n* (in the houses) is the grammatically correct form. Likewise, the singular form *Haus* is the same for all cases except for the genitive: *Haus-es*.

Now, with regards to the design of a lexical decision task, two groups of target words were distinguished: Words ending in *–n*, and words ending in *-s* (N-Group and S-Group). Furthermore, these two groups were then split up according to the 'grammatical load' of the suffixes, which renders three subgroups each (N1, N2-PL, N3-PL-DAT, S1, S2-SG-GEN, S3-PL-GEN) as depicted in table 1.

Table 1: Dative and genitive groups with grammatical load indicated by colors.

	Grammatical load	Example
Group N1	-n part of stem (low)	*Zahn* (tooth)
Group N2-PL	-n denoting plural for all cases (medium)	*Rabe-n* (ravens)
Group N3-PL-DAT	-n as distinctive dative plural marker (high)	*Stiefel-n* (boots.DAT)
Group S1	-s part of stem (low)	*Gleis* (platform/track)
Group S2-SG-GEN	-s as genitive singular marker (medium)	*Pferde-s* (horse.GEN)
Group S3-PL-GEN	-s as genitive singular and plural marker for all cases (high)	*Zoo-s* (zoos)

*Umlaut was avoided, except for *Ästen* (branches)

As can be seen in Table 1, the groups are put together according to different functions of the final *-n* and *-s*. They might not have any grammatical function (groups N1 and S1), they can have one specific function, namely denoting the plural (group N2-PL) or the genitive singular (S2-SG-GEN), or they can represent two different grammatical functions – both plural and case marking – as in groups N3-PL-DAT and S3-PL-GEN.

In order to also control for potential frequency effects, the WEBCELEX[2] database was used to select 20 target words for each of the 6 groups. These 120 target words were matched for *surface frequency* (ranging from 20-1 per ~5 million) and length in letters (ranging from 3-10 letters per word). Additionally, data on other frequency measures such as *stem frequency*, *type frequency*, *family size* and *family frequency*[3] was also included.

Although we include frequency as a potential confound in the model, it will be shown in the LDT that the more grammatical load an affix carries, the harder it will be for participants to decide whether the inflected noun is a correct German word form or not (reflected in longer response latencies), irrespective of token frequency effects.

3. Lexical Decision Experiment

3.1 Methods

Participants. A lexical decision task was performed with 26 participants volunteering to participate in the study, all of them native speakers of German with a mean age of ~27 (14 females, 12 males).

Materials. The above mentioned 120 target words – split up into 6 groups (N1-S3) – were selected from the WEBCELEX database and matched for surface frequency and length in letters within groups. Additionally, 120 random filler words were selected from WEBCELEX, as well as 240 non-words of which 120 were produced by manually changing two or three letters of the stem (of other words in WEBCELEX), and 120 by changing potential affixes. This way, subjects were prevented from relying solely on recognition of stems for their lexical decision. All non-words adhered to the phonotactic rules of German. All filler words and non-words were chosen to reduce possible priming effects with regard to the target words. Overall the number of words and non-words added up to 480 items.

Items were presented by using the *SuperLab 4.5.2* stimulus presentation software (Abboud et al., 2011). To present the stimuli, the item list was split up into three blocks with 160 items each, which all contained roughly the same ratio of target words, filler words and non-words. Items were presented as black *Tahoma* letters in font size 20 against a light turquoise background. They were preceded by a black fixation point in the center of the screen for 500ms before stimulus onset. There was no time limit for

[2] Online: http://celex.mpi.nl/

[3] *Surface frequency* denotes the token frequency of a word form (such as *table*) (Schreuder & Baayen, 1997: 119). *Stem frequency* (Schreuder & Baayen, 1997: 120) is derived by cumulating frequencies of *inflectional variants* of a word, which have also been shown to play a role in reaction time experiments (Nagy et al., 1989; Alegre & Gordon, 1999). Moreover, the *family size* of a word is the stem frequency + the number of derived words (e.g. *health/health-y*) and the number of compounds (*e.g. table/tablecloth*) (Schreuder & Baayen, 1997; Bertram et al., 2000). Finally, the *family frequency* of a word is the sum of frequencies of all the forms belonging to the same morphological family.

Besides this class of *token frequencies*, which are used to predict RTs for lexical entries and lemmas of words, there is the concept of *type frequencies*, too, which captures the number of different words inflected with a particular marker (e.g. the number of verbs which are inflected with regular *-ed* versus the number of irregular verbs) (Bybee, 2007; Marcus et al., 1995: 212).

responses. Participants responded to stimuli by using a *Cedrus* response pad (model RB-730) with green and red buttons for word and non-word decisions.

For statistical analyses and data plotting, the software *R* (R Development Core Team, 2010) was used. Additionally, the software packages *lme4* (Bates & Maechler, 2010) and *languageR* (Baayen, 2010; cf. Baayen, 2008) as well as *ggplot2* (Wickham & Chang, 2012) were used to construct linear mixed-effects models as well as to get boxplots.

Procedure. In the instructions participants were told to decide as quickly and accurately as possible whether the presented items are German words or not. They were explicitly told that forms with plural and case inflections can be part of the stimulus set. Then they were presented with a test trial containing 8 words and 8 non-words. Both dative and genitive marked words were represented in this set of test items. In the test trial items remained on the screen until the participant had pressed the correct button. The instructor remained in the room during the test trail and participants were able to ask questions. After that the instructor left the room and participants were presented with the three blocks of 160 items each (with one minute pauses in between). The testing took 15-20 minutes.

After finishing the main experiment, participants were presented with a questionnaire to clarify 1) whether they had guessed what the exact purpose of the experiment is; 2) whether they had issues with specific items; 3) whether they had used any specific strategy to decide on words with dative and genitive marking. Participants could use the keyboard to type their answers, but they were also told that they can just type “no” if they did not want to answer the questions.

3.2 Results

A pre-analysis of the data revealed that 4 of the 26 participants had to be excluded from the dataset because they had guessed the purpose of the experiment. Also, three of the items[4] had to be excluded because their per item error rate exceeded 50%. The error rates per subject ranged from 1.6% to 21%. No further subjects were excluded. This left 22 subjects and 117 items to be analyzed. Furthermore, RTs were cleaned by excluding all RTs of less than 300ms for reasons of lower processing bounds (Baayen, 2008: 243). Also, all RTs longer than 3000ms were excluded because both inspection of *quantile-quantile plots* (Baayen, 2008: 243) as well as considering 2-3 standard deviations from the overall mean (mean: 959ms; SD: 934ms) as a cut-off point suggested that 3000ms are a realistic upper bound for RTs. Moreover, for the analysis of reaction times all incorrect responses were excluded from the sample. These cleaning procedures caused an additional data loss of ~8%.

In the following, the RTs for the N-Groups and S-Groups are analyzed separately. Plotting the subgroups and logarithmically transformed RTs for each group reveals that there are indeed differences in mean reaction times (see figure 1a and 1b).

In order to check whether the results are significant if groups are compared by subjects and items, a *linear mixed-effects model* (Baayen, 2008.; Baayen et al., 2008) with RTs (logarithmically transformed) as *dependent variable* and subgroups as *predictor variables* as well as subjects and items as *crossed random effects* was used. For tests of significance, *Markov Chain Monte Carlo*-estimated p-values are presented (Baayen et al., 2008, p. 397-398).

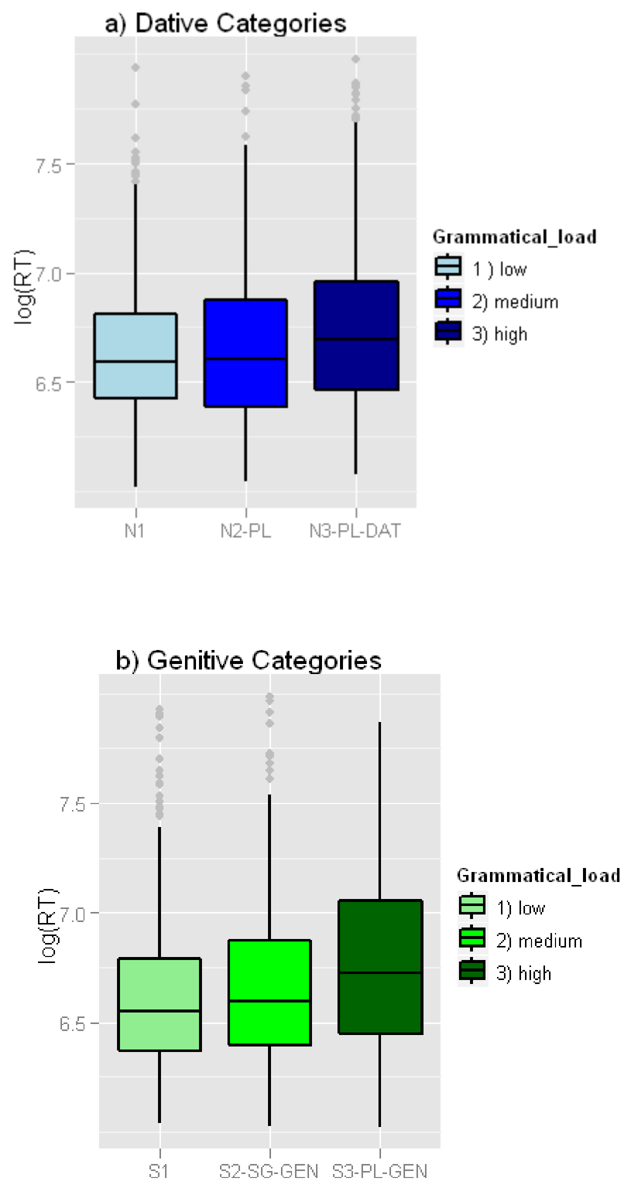

Figure 1: Boxplots for log(RTs), inflectional categories (x-axes), and grammatical load indicated by color.

Model validation was performed by a) checking homoscedasticity and normality for plots of residuals versus fitted values, b) by means of likelihood ratio tests of the test model against a null model (no fixed effects).

[4] *Fries* (frieze), *Schahs* (shahs), *Gemischen* (mixtures.DAT)

This model reveals that words in N1 are processed with significantly lower response latencies than words in the grammatically loaded N3-PL-DAT group ($p < 0.001$). The same holds for S1 and S3-PL-GEN ($p < 0.001$). Moreover, words in the N1-PL are still associated with significantly lower RTs compared to N1-PL-DAT ($p < 0.05$). Likewise, S1-GEN-SG words that only mark case elicit shorter response latencies than S1-PL-GEN words, which potentially mark case and/or plural ($p < 0.05$). When comparing the unmarked N1 and S1 groups with the groups marked for plural (N2-PL) and genitive case (S2-SG-GEN) only, the effects also go in the right direction (i.e. longer mean RTs for the marked groups), however, these effects are non-significant ($p > 0.05$).

Finally, using ANOVAS for comparisons of both the N-Group and S-Group models with null-models (without subgroups as fixed effects) reveals that adding subgroups as predictors significantly enhances model fits (N-Group: $\chi^2(2) = 11.65$, $p < 0.01$; S-Group: $\chi^2(2) = 13.65$, $p < 0.01$). Overall, this suggests that the central effects obtained in the LDT generalize over subjects and items.

Now, in order to contrast these results with the predictive power of frequency effects on RTs, another mixed-effects model was designed, this time with *surface frequency*, *stem frequency*, *type frequency*, *family frequency* and *family size* as predictor variables as well as subjects and items as random effects. In this model *type frequency* is the only significant predictor of RTs ($p_{typeFreq} < 0.05$) the other frequency measures are not predictive ($p_{surfaceFreq} = 0.23$; $p_{stemFreq} = 0.52$; $p_{famFreq} = 0.28$; $p_{famSize} = 0.43$)[5]. Also, an ANOVA comparing the mixed-effects model and a null-model revealed that there is only marginal enhance in predictive power if the frequency measures are added as fixed effects ($\chi^2(5) = 10.07$, $p = 0.073$). Moreover, since *family size* and *family frequency* are strongly correlated ($r = -0.82$), two more models with only family size and family frequency as predictor variables were fitted to avoid these predictors from cancelling each other out. However, even in these models both family size ($p = 0.7$) and family frequency ($p = 0.26$) did not predict the patterns in RTs.

Finally, it should be noted that all the linear mixed-effects models presented in this section are more or less “stressed” for longer response latencies in the RT data. This follows logically from the fact that RT distributions are somewhat skewed, exhibiting longer right tails. However, as will be argued in the following section, it is exactly the occurrence of non-normally prolonged response latencies that is interesting for the overall interpretation of the data.

4. Discussion

The results reported for the lexical decision task suggest that there are systematic differences between nouns for which the *-n* and *-s* suffixes are grammatically meaningless (N1 and S1 subgroups in table 1) and nouns which are grammatically loaded (N3-PL-DAT, S3-PL-GEN). Moreover, subgroups which are inflected for plural or case only (subgroups S2-SG-GEN and N2-PL) lie somewhere in between the unmarked nouns and the heavily marked nouns in terms of reaction times. Interestingly, the observed patterns of reaction times per subgroup are not predicted by measures of token frequency. Token frequencies could not be shown to be significant predictors of RTs in post-hoc regression analyses.

The only marginally predictive frequency measure is *type frequency*. Now, it is important to be aware of the fact that type frequencies are tied with subgroups N1-S3 since they reflect the ‘inflectional status’ of a word, which is in turn the grouping factor for further divisions of the N-Group and S-Group. For example, all the words in N1 have a type frequency of 15926/35315 (45% of all the nouns in WEBCELEX), whereas all the words in N3 have a type frequency of 3140/35315 (8.9%). Likewise, all the nouns in N1 share the inflectional status of being unmarked for case or plural and all the nouns in N3 share the inflectional status of being marked for plural and case. These were basically the search criteria for finding appropriate nouns in WEBCELEX. Hence, type frequency and subgroup membership are two sides of the same coin.

At this point the question arises what actually causes the longer response latencies. According to dual route accounts there are two possible explanations: a) Differences in token frequencies have an impact via the direct lexical access route – this has been ruled out by controlling for surface frequency in the experiment and by including other measures of token frequencies in a post-hoc multiple regression model; or b) The differences in RTs stem from parsing difficulty for complex morphological structures *within* the words (see parsing example in figure 2).

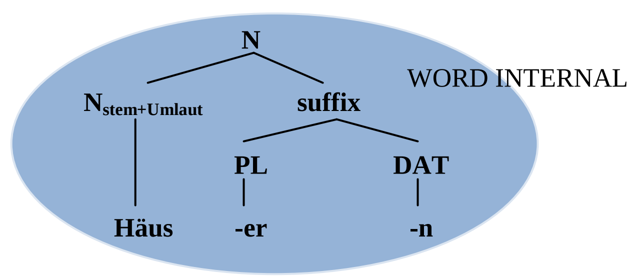

Figure 2: Potential word internal structure for the morphologically complex noun *Häusern* (houses.DAT) with both plural and dative marking.

However, a third possibility is that the differences in RTs are due to the additional grammatical and conceptual load that these suffixes carry. This means, rather than analyzing structures within the word, it would be more interesting to analyze the *context* these words are typically embedded in. See, for example, a typical sentence involving the noun *Häusern* in German (figure 3).

[5] Of course, we do not expect surface frequency to be predictive anyway, because items were matched within groups. This is not to say, however, that the other token frequency measures cannot be predictive, since surface frequency and other types of frequencies can differ significantly.

This figure illustrates the grammatical relationships between the word internal and word external structure. The dative marking is triggered by a preposition *hinter* (behind) (i.e. lexical case). Moreover, the plural form needs to agree with the DAT.PL of the article *die*.SG, i.e. *den*.DAT.PL. Hence the word *Häusern* is embedded into a construction that involves a preposition and a case-marked article. We could think of more such examples with other prepositions (e.g. *auf* (on top of), *in* (in), *mit* (with)).

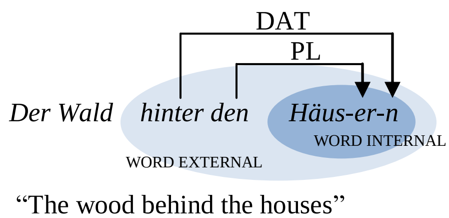

Figure 3: Grammatical relationships between elements of a sentence involving dative marking.

Crucially, note that the *type frequency* of this dative plural marker, i.e. the range of words it is applied to, hinges upon the productivity of such prepositional constructions (plus the productivity of dative forms in other contexts). This would suggest that increased processing difficulty in the LDT for grammatically loaded words stems from the strength of embedding into common or uncommon constructions.

Of course, there needs to be further research with and beyond LDTs to further elaborate this hypothesis. However, hints suggesting that this explanation is along the right lines can be found in the questionnaire.

4.1 Questionnaire

When the first participant to be tested came across the German word *Messers* (knife's) in the trial set, he kept pressing the 'non-word' button several times, although this is a grammatically correct form and the item kept occurring on the screen. When the instructor noted that this is a genitive form of the word *Messer*, the participant said: "... *auf Messers Schneide*!" A German idiom directly translated as: "on knife's blade", meaning: "to be on a knife-edge".

Evaluating the post-test questionnaire revealed that this spontaneous associative reaction might not have been a single coincidence. When asked (question 2) whether they had particular problems with specific items, 10 (45%) of the participants answered "no", 6 (27%) of the participants answered that they had problems with either dative, genitive or plural forms, and the rest (28%) named non-words and potential foreign words as problematic. Most intriguingly, when subjects were more specifically asked (question 3) whether they had specific problems with case-marked words (by giving them some examples of the target set) and whether they used a "trick" to decide upon these words, 13 (52%) answered with "no", 6 (24%) answered that they had imagined the correct articles to take a decision, and 5 (20%) had even used phrases like "*die Spitze des Doms*" (Eng. the cathedral's spire) or prepositional phrases "*wegen des Kochs*" (Eng. because of the cook) to take their decision, and one participant noted that he had focused on the pronunciation of words and potential umlaut.

In order to test whether the strategies named here might have prolonged reaction times, participants were post-hoc divided into two groups: one group (*no-context group*) for subjects that had negatively answered questions 2 and 3 (or who had named other difficulties like non-words and foreign words), and another group for subjects that had answered affirmative and noted that they used context related strategies to take lexical decisions (*context group*). Interestingly, for these two groups the mean RTs for S3 and N3 taken together differ: For the *context group* the mean RTs for words in S3 and N3 is slightly higher (956ms) than for the *non-context group* (939ms), although this difference is not significant (p=0.33).

However, the fact that 12 (55%) of the participants either had problems with dative and genitive markers or used "minimal phrases" as disambiguation strategy suggests that this is at least partly the reason for prolonged response latencies. Note that the rest of the participants (10, 45%) did not necessarily use some other strategy or no strategy at all. Rather, participants could just type "no" if they did not want to bother with the questionnaire in the first place. Overall, the insights from the questionnaire suggest that there are systematic reasons for prolonged response latencies, namely, whether forms are more or less embedded into usage-based units.

5. Conclusion

In the past, lexical decision tasks have been invoked to find out whether certain lexical items are processed as a whole or decomposed into *stem+affix*. In this context, it has been argued that for entries stored as whole units in the lexicon there should be surface frequency or other token-related frequency effects observable, whereas for morphologically complex and regular items symbolic rules will be applied which are not sensitive to frequency effects (Marcus et al., 1995; Clahsen, 1999; Pinker & Ullman, 2002). However, the results reported in this paper suggest an alternative to this binary distinction.

First of all, it has been shown that token frequencies are not a significant predictor when it comes to morphologically complex nouns in German, whereas grammatical load and type frequencies do still correctly predict longer reaction times for these forms. Thus, instead of trying to explain response latencies by analyzing morphological structures within the lexical items, this study suggests that the relevant factor is the embedding of these items in more or less frequent phrases. This is in line with accounts arguing that statistical learning and frequency effects are not only relevant for "lexical entries" but also for whole constructions (Ellis & O'Donnell, in preparation).

In conclusion, there are measurable processing differences between grammatically marked and unmarked nouns in German. Hence, it is correct, on principal, to distinguish between words that are perceived as “basic” or “default” and words which are perceived as grammatically complex. However, this does not necessarily entail that such morphologically complex forms are composed out of simpler units by means of symbolic rule manipulation. Rather, such forms carrying more ‘grammatical load’ are more likely to be associated with whole phrases and sentences even in isolation. And this embedding in redundant and disambiguating structures is what makes them belong to the grammatical rather than the lexical domain in the first place.

References

Abboud, H., Heller, K., Matsak, E., Schultz, H., & Zeitlin, V. (2011). SuperLab. Stimulus presentation software. San Pedro, California: Cedrus Corporation.

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

Baayen, H. R. (2010). *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics”*. Retrieved from http://CRAN.R-project.org/ package= languageR.

Baayen, H. R. (2008). *Analyzing linguistic data: A practical introduction using R*. Cambridge: Cambridge University Press.

Baayen, H. R., 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.

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes. Retrieved from http://cran.r-project.org/package=lme4

Bertram, R., Baayen, R. H., & Schreuder, R. (2000). Effects of family size for complex words, *42*, 390–405.

Bybee, J. (2007). *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Clahsen, H. (1999). Lexical entries and rules of language: a multidisciplinary study of German inflection. *Behavioral and Brain Sciencesrain sciences*, *22*(6), 991–1060. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11301574

Clahsen, Harald, Eisenbeiss, S., & Sonnen-Henning, I. (1997). Morphological structure and the processing of inflected words. *Theoretical Linguistics*, *23*, 201–249.

Ellis, N. C., & O’Donnell, M. B. (2012). Statistical construction learning. In J. N. Williams & P. Rebuschat (Eds.), *Statistical learning and language acquisition*. Berlin: Mouton de Gruyter.

Engel, U. (1991). *Deutsche Grammatik*. Heidelberg: Julius Groos.

Griesbach, H. (1986). *Neue deutsche Grammatik*. Berlin: Langenscheidt.

Hahn, U., & Nakisa, R. C. (2000). German inflection: single route or dual route? *Cognitive Psychology*, *41*(4), 313–60. doi:10.1006/cogp.2000.0737

Kempe, V., & MacWhinney, B. (1998). The acquisition of case marking by adult learners of Russian and German. *Studies in Second Language Acquisition*, *20*(4), 543–587. doi:10.1017/S0272263198004045

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*, 189–256.

Marslen-Wilson, W. D., & Tyler, L. K. (2007). Morphology, language and the brain: The decompositional substrate for language comprehension. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *362*, 823–836. doi:10.1098/rstb.2007.2091

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past-tense inflections: What does the evidence rule out. *Trends in Cognitive Sciences*, *6*(11), 456–463.

Nagy, W., Anderson, R. C., Schommer, M., Scott, J. A., Stallman, C., Nagy, W., Anderson, R. C., et al. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*(3), 262–282.

Pinker, S., & Ullman, M. T. (2002). The past and the future of the past tense. *TRENDS in Cognitive Science*, *6*(11), 456–463.

Sonnenstuhl, I., & Huth, A. (2002). Processing and representation of German -n plurals: A dual mechanism approach, *81*, 276–290.

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

Wickham, H., & Chang, W. (2012). ggplot2: An implementation of the grammar of graphics. Retrieved from http://cran.r-project.org/web/packages/ggplot2/index .html