Towards exploring the specific influences of wordform frequency, lemma frequency and OLD20 on visual word recognition and reading aloud

*Lara Kresse, Stefan Kirschner, Stefanie Dipper, Eva Belke**

It is well established that the speed and accuracy of visual word recognition and reading aloud are influenced by a multitude of variables, including the frequency of the word, its orthographic and phonological neighbourhood, the consistency between its written and spoken wordforms, etc. Among these variables, word frequency and the structure of the wider orthographic neighbourhood in terms of a word's *N* or OLD20 (the average Levenshtein distance of a word's 20 nearest Levenshtein neighbours) appear to be the variables that have been studied most intensively (see Yarkoni, Balota, & Yap, 2008; for review, see Balota, Yap, & Cortese, 2006; Rastle, 2007). To our knowledge, most studies of visual word processing concerning the role of word frequency rely on wordform frequencies as a measure of word frequency (note, however, that this is hardly ever stated explicitly). While this may appear reasonable given that the object of study is the recognition of individual wordforms, using wordform frequencies may prematurely exclude important information about the relatedness of wordforms through their shared lemma (cf. Ford, Marslen-Wilson, & Davis, 2003; see Crepaldi, Rastle, Coltheart, & Nickels, 2010, for a proposal on how

*Ruhr-Universität Bochum

such relations between wordforms with shared lemmas could be conceived of in current models of visual word recognition):

- Firstly, lemmas may differ with respect to the number of wordforms they are associated with (henceforth referred to as wordform class).
- Secondly, individual members of the same wordform class may be more or less frequent instances of their shared lemma. Figure 2.1 plots, for a set of wordforms pertaining to 500 frequent lemmas, the wordform frequencies (in light grey) and the frequencies of their associated lemmas above them (in dark grey) (see below for a more detailed description of the German newspaper corpus the wordforms had been extracted from). Overall the wordform frequencies associated with all lemmas are quite low, suggesting that many different wordforms contribute to the overall lemma frequency.
- Thirdly, apart from some exceptions, wordforms with shared lemmas are usually orthographically and morphologically similar. A straightforward prediction would be that, on average, wordforms that belong to a large wordform class should be associated with lower OLD20 scores than wordforms from smaller wordform classes.

In the present research, we intended to explore the specific influences of wordform frequency, lemma frequency, and OLD20 on visual word processing. We worked with a corpus consisting of 7 volumes of the Neue Züricher Zeitung (1993–1999; 175 million tokens). The corpus has been annotated with part-of-speech tags and lemmatized by the RFTagger (Schmid & Laws, 2008). We extracted all common nouns (wordforms) from the corpus. Given that word length is an important variable in visual word processing, we restricted the length of the wordforms to 6 to 10 characters. We selected relatively long wordforms in order to exclude extremely-frequent outliers, which are usually very short. Next, we selected a random subset of 10,000 entries and assembled eight groups of

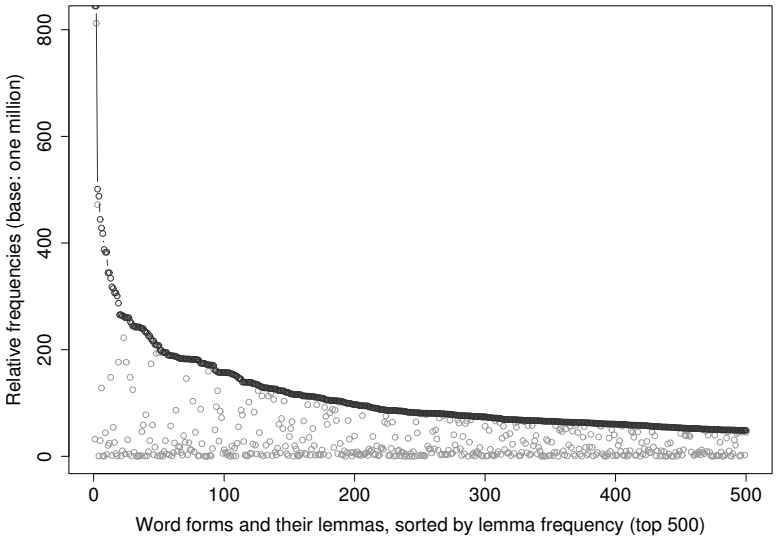


Figure 2.1: Frequencies per million of 6- to 10-letter-long wordforms (in light grey) and their associated lemmas (in dark grey). Only the data pertaining to the 500 most frequent lemmas is shown. The data is sorted by lemma frequencies, and the frequencies of a lemma and its wordforms are vertically aligned.

wordforms by factorially varying the selection specifications with respect to three criteria:

(i) high or low lemma frequencies

(ii) high or low wordform frequencies

(iii) high or low OLD20-scores

To this end, we first determined the lemmas pertaining to each of the 10,000 wordforms, established the quartile sections of the lemma frequencies, and selected the wordforms in the first and last quartile sections. Within each of these quartile sections, we

then selected the first and last quartile sections with respect to the wordform frequency of the entries. The resulting four sets of wordforms were then all split in quartiles with respect to their OLD20 scores, which had been computed initially with reference to the full set of 10,000 wordforms. On average, this yielded 161 entries per group.

Our first goal was to study the makeup of the eight sets with respect to their lexical and morphological properties. The results are summarized in Table 2.1, which also provides examples of wordforms from each set. As expected based on previous findings (see Yarkoni et al., 2008), wordforms with low OLD20 scores tended to be shorter than wordforms with high OLD20 scores [criterion (iii); low: average length of 6.8–7.4 characters (see columns 2, 4, 6, 8 in Table 2.1); high: 9.0–9.5 characters (columns 1, 3, 5, 7 in Table 2.1)]. Furthermore, wordforms with low OLD20 scores were often inflected for case and/or number, suggesting that, as predicted, they were members of a larger wordform class. Comparing sets with high vs. low wordform frequencies (criterion (ii)), we found a clear difference with regard to morphological properties: Sets with low wordform frequencies contained a considerable amount of deverbal and deadjectival nominalizations [low: 23 %–70 % (columns 3, 4, 7, 8 in Table 2.1); high: 0 %–8 % (columns 1, 2, 5, 6 in Table 2.1)]; examples of such nominalizations are *Relevantes* ('relevant points') or *Besprühen* ('spraying'). Compounds occurred rarely in general, most likely due to the restriction of the overall length of the wordforms to ten letters. With regard to criterion (i) — lemma frequencies — no clear distinctions between both extremes stood out in the analysis of the dimensions listed in Table 2.1 (low: columns 5–8; high: columns 1–4), possibly because the restriction to 6–10 letter wordforms constrained the lemma frequencies to a too narrow range.

Table 2.1: Overview of the properties of the eight wordform sets. Set labels indicate the lemma frequency class in the first letter (H = high, L = low), the wordform frequency class in the second letter, and the OLD20 class in the last letter. The first three rows present the lower end of the forth quartile section and the upper end of the first quartile section for each of these dimensions. Frequency counts for lemma frequencies (LF) and wordform frequencies (WFF) are given as number of occurrences per 1 million. The next three lines provide the number of entries per set, the average word length per set (with SD) and the average OLD20 scores of each set (with SD). The next three rows present the proportions of deverbal and deadjectival nominalizations (NVinf/NAdj), of compounds and of all other nouns. In the last row, we provide examples of the items included in the individual sets.

	Set							
	1 HHH	2 HHL	3 HLH	4 HLL	5 LHH	6 LHL	7 LLH	8 LLL
LF	> 4.33				< 1.28			
WFF	> 9.44		< 0.70		> 0.05		< 0.01	
OLD20	> 3.25	< 2.30	> 3.25	< 2.30	> 3.25	< 2.30	> 3.25	< 2.30
# Entries	115	190	193	139	114	215	228	92
# Characters	9.0 ± 1.0	6.9 ± 1.2	9.2 ± 0.8	7.4 ± 1.2	9.2 ± 0.8	6.8 ± 0.9	9.5 ± 0.7	7.3 ± 1.0
OLD20	3.7 ± 0.3	1.9 ± 0.2	3.8 ± 0.4	1.9 ± 0.2	3.6 ± 0.3	1.9 ± 0.2	3.7 ± 0.4	2.0 ± 0.2
NVinf/NAdj	.00	.03	.23	.55	.04	.08	.70	.61
Compound	.09	.00	.03	.00	.08	.00	.00	.00
Other	.91	.97	.74	.45	.88	.92	.30	.39
Examples	Nachmittag; Polizisten	Geschäfte; Projekten	Einzelheit; Teppichs	Paletten; Schattens	Tapezierer; Absprünge	Käuzchen; Noppen	Billiarde; Frotzelei	Schöpfens; Stoppel

Next, we prepared two experiments to have a group of participants perform lexical decision and word naming tasks on subsets of wordforms from the eight sets. The subsets were matched as closely as possible for number of morphemes, rated familiarity and rated concreteness (see Table 2.2 for details). However, OLD20 turned out to be unavoidably confounded with word length, preventing us from matching all eight sets for word length (see Table 2.2). In total, 176 wordforms were tested. For the lexical decision task, we created pseudoword partners to all wordforms, using the German version of the Wuggy pseudoword generator (Keuleers & Brysbaert, 2010). This software generates pseudowords that match real words in terms of their phonotactic properties (number of syllables, subsyllabic structure, phoneme transition frequencies), their number of characters, and their OLD20 scores. Given that the experimental stimuli consisted of nouns only and that the first letters of nouns are capitalized in German orthography, we employed a case-sensitive version of Wuggy, which ensured that the initial syllables of the pseudowords were selected from a pool of first syllables of capitalized words (nouns) as well. Examples of the resulting pseudowords are *Bisivisten* and *Hästeln* (as counterparts of the real wordforms *Polizisten* (Set HHH) and *Misteln* (Set LLL), respectively).

In each task, the stimulus sets were tested in a blocked fashion. Each participant was tested on a different order of sets but worked through the same order of sets in both tasks. The order of stimuli within sets was fully randomized within and across participants. In both experiments, participants saw a fixation cross for 500 ms, followed by a blank screen for 50 ms and the target word, which was shown until the participants' response was registered or until 1200 ms had elapsed. 760 ms later, the next trial was initiated.

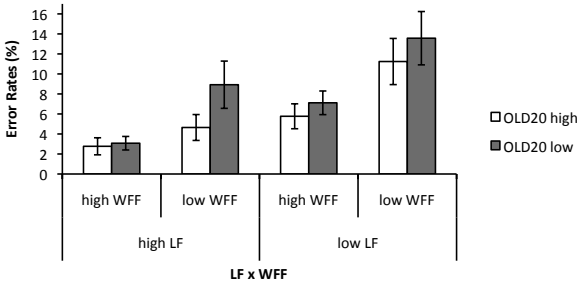
The order of completing the lexical decision and word naming tasks was counterbalanced across participants with half of the participants completing the word naming task first and the other half completing the lexical decision task first. 26 undergraduate students participated in the experiment.

Table 2.2: Overview of the properties of the stimulus sets used in the lexical decision and word naming tasks. Frequency counts are given as number of occurrences per 1 million. The subsets were matched as closely as possible for number of morphemes, rated familiarity, rated concreteness, and, less successfully, word length.

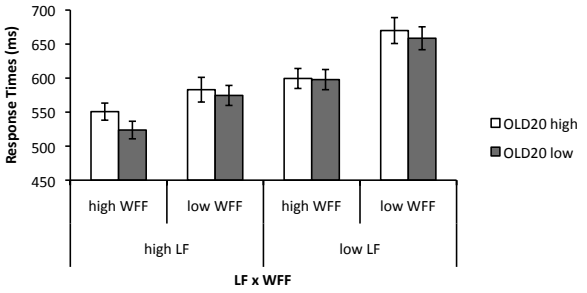
Set	N	Avg. LF	Avg. WFF	Avg. OLD20	Avg. # Characters	Avg. # Morphemes	Median familiarity ratings	Median concreteness ratings
1 HHH	25	42.64	20.87	3.56	9.12	2.24	1.00	2.00
2 HHL	25	99.62	21.30	2.01	7.24	2.24	1.00	2.00
3 HLH	25	18.34	0.33	3.64	9.04	2.32	2.00	2.00
4 HLL	25	17.01	0.35	2.01	7.28	2.12	2.00	2.00
5 LHH	25	0.09	0.09	3.59	9.00	2.28	2.00	2.50
6 LHL	20	0.01	0.09	2.01	7.25	1.90	2.00	1.50
7 LLH	14	0.02	< 0.01	3.73	9.14	2.21	2.50	2.75
8 LLL	17	0.01	< 0.01	1.99	7.12	2.41	3.00	2.00

The data analyses were restricted to the critical stimuli featuring in both tasks, excluding "no" responses in the lexical decision task. An initial inspection of the error rates in the lexical decision task showed that three wordforms were misclassified as non-words by at least 50 % of the participants: *Einnistens* (Ein-nisten-s, in-settle-GEN.SG, 'of settling in'; Set LLH), *Havaristen* (Havarist-en, disabled_vessel-GEN.SG;DAT.SG;PL, 'disabled vessel(s)'; Set LHH), *Amnestien* (Amnestie-n, amnesty-NOM.PL;ACC.PL, 'amnesties'; Set HLH). After exclusion of the data pertaining to these three items from both tasks, participants' individual error rates in the lexical decision task ranged from 1.2 % to 20.2 % (M = 6.8 %, SD = 3.9 %). In the word naming task, individual error rates ranged from 0.6 % to 16.7 % (M = 4.0 %, SD = 3.7 %) with an additional 0.3 % voicekey errors across all participants. Error rates were submitted to analyses of variance by participants (F_1) and by items (F_2) with lemma frequency (LF), wordform frequency (WFF) and OLD20 as within-participants and between-items variables. Prior to parallel analyses of response times, all valid response times exceeding 3.5 standard deviations of a participants' mean in the lexical decision task were excluded (1.5 % of all valid data), leaving 4138 and 4272 data points for further analyses in the lexical decision task and the word naming task, respectively.

As shown in Figures 2.2 and 2.3, error rates and response times in both tasks were affected by wordform frequency and lemma frequency, with more accurate and faster response times for words with high lemma and/or wordform frequencies than for words with low lemma and/or wordform frequencies. In the lexical decision task, words with low OLD20 scores yielded higher error rates but shorter response times than words with high OLD20 scores. In the word naming task, words with low OLD20 scores were named faster and more accurately than words with high OLD20 scores. Analyses of variance (ANOVAs) by participants and by items of participants' error rates and response times yielded significant effects of lemma frequency and wordform frequency in both tasks (see Tables 2.3 and 2.4). In the analyses of lexical decision times, the



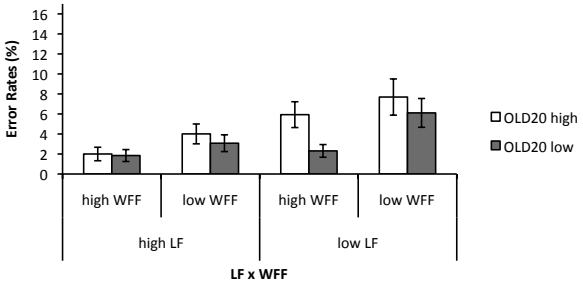
(a)



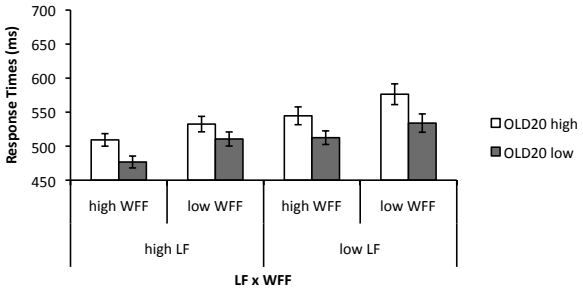
(b)

Figure 2.2: Error rates (a, in %) and response times (b, in ms) in the lexical decision task, broken down by lemma frequency (LF), wordform frequency (WFF) and OLD20. Error bars represent one standard error.

interaction of wordform and lemma frequency approached significance with stronger effects of lemma frequency on low-frequency than on high-frequency wordforms (see Figure 2.2b). In lexical decision, OLD20 had overall weaker effects than lemma or wordform frequency (see Table 2.3). However, it significantly impacted on the word naming times (see Table 2.4). As outlined above, participants were faster to name words with a low OLD20 score than words with a high OLD20 score. The finding that OLD20 affected word naming times but not lexical decision times suggests that its effect



(a)



(b)

Figure 2.3: Error rates (a, in %) and response times (b, in ms) in the word naming task, broken down by lemma frequency (LF), wordform frequency (WFF) and OLD20. Error bars represent one standard error.

can be localized at the phonological encoding stage of the word naming task where words with a low OLD20 score will be easier to encode than words with a high OLD20 score.

Given that we failed to fully match the eight sets on the control variables listed above (length, number of morphemes, median rated familiarity scores and median rated concreteness scores, Baayen, 2010), we entered these control variables and the three target variables (OLD20, log-transformed lemma frequencies and log-transformed wordform frequencies of all wordforms) into forward

Table 2.3: Results of analyses of variance by participants (F_1) and by items (F_2) of the error rates and response times in the lexical decision task. (*** $p < .001$, ** $p < .01$, * $p < .05$, $^{(*)}$ $p < .1$).

	Error rates (%)	
	F_1 (1,25)	F_2 (1,165)
LF	17.43 ***	15.41 ***
WFF	12.36 **	17.81 **
OLD20	4.87 *	3.15 $^{(*)}$
	Response Times (ms)	
	F_1 (1,25)	F_2 (1,165)
LF	104.86 ***	140.07 ***
WFF	52.04 **	78.26 **
OLD20	2.99 $^{(*)}$	2.53 $^{\text{n.s.}}$
LF $\times$ WFF	6.56 *	3.76 $^{(*)}$

Table 2.4: Results of analyses of variance by participants (F_1) and by items (F_2) of the error rates and response times in the word naming task. (*** $p < .001$, ** $p < .01$, * $p < .05$, $^{(*)}$ $p < .1$).

	Error rates (%)	
	F_1 (1,25)	F_2 (1,165)
LF	22.10 **	7.52 **
WFF	8.62 ***	4.72 *
OLD20	7.67 **	2.41 $^{\text{n.s.}}$
	Response Times (ms)	
	F_1 (1,25)	F_2 (1,165)
LF	87.08 ***	55.56 ***
WFF	48.08 ***	33.75 ***
OLD20	85.02 ***	44.96 ***

regression analyses of all valid response times (with $p_{in} = .05$). In the analyses of lexical decision times, rated familiarity was included first, followed by log lemma frequency, wordform length, and log wordform frequency (see Table 2.5). The final model accounted for

9.0 % of the variance. In parallel analyses of the word naming times, wordform length was entered first, followed by rated familiarity, log wordform frequency, OLD20, and log lemma frequency (see Table 2.5), eventually accounting for 8.4 % of the variance. Overall, the results from the regression analyses correspond to the findings from the ANOVAs of the response times in each task, which had yielded significant effects of wordform frequency and lemma frequency on lexical decision and word naming times and of OLD20 on word naming times. The regression analyses indicate that, as expected, there were confounds of OLD20 and word length and of log lemma frequency and rated familiarity.

Table 2.5: Results of the forward regression analyses of all valid response times in the lexical decision and the word naming tasks. The predictor variables are listed in the order they were entered in the regression analyses along with the increase in R^2 they accounted for.
(*** $p < .001$, ** $p < .01$, * $p < .05$, $^{(*)}$ $p < .1$).

	B	SE	β	t	R^2 change
	Lexical Decision				
Median Familiarity	27.502	2.711	.177	10.146 ***	.066
logLF	−3.249	0.816	−.110	−3.982 ***	.017
Length (# chars)	7.301	1.477	.074	4.945 ***	.005
logWFF	−1.543	0.719	−.059	−2.145 *	.001
	Word Naming				
Length (# chars)	8.397	1.304	.134	6.438 ***	.032
Median Familiarity	12.52	1.7	.128	7.363 ***	.036
logWFF	−1.154	0.434	−.07	−2.659 **	.010
OLD20	9.879	2.078	.099	4.754 ***	.005
logLF	−1.063	0.499	−.056	−2.13 *	.001

In summary, our results demonstrate that lemma frequency has an effect over and above wordform frequency. From a methodical point of view, this suggests that when compiling material for studies on visual word processing, both lemma frequencies and wordform frequencies must be considered. With respect to cur-

rent accounts of visual word recognition and reading aloud, our data suggest that there are independent levels of representation of lemma and wordform representations, as has been proposed, for instance, by Crepaldi et al. (2010). While Crepaldi et al. (2010) postulate interactive connections between lemmas and their various wordforms, the data reported in the present study indicate that the interaction of these processing levels is rather weak – the interaction of wordform frequency and lemma frequency was marginally significant in the analysis of lexical decision times but did not reach significance in the analyses of word naming times. In future work, we seek to establish more carefully controlled item sets in order to assess whether lemma frequency and wordform frequency effects interact or not. If they do not, as is suggested by the present findings, this will have important implications for current accounts of visual word processing, such as the one proposed by Crepaldi and colleagues.

Contact: Eva Belke `<belke@linguistics.rub.de>`
Stefanie Dipper `<dipper@linguistics.rub.de>`

References

Baayen, R. (2010). A real experiment is a factorial experiment? *The Mental Lexicon*, *5*(1), 149–157.

Balota, D., Yap, M., & Cortese, M. (2006). Visual word recognition: The journey from features to meaning (a travel update). In M. Traxler & M. Gernsbacher (Eds.), *Handbook of psycholinguistics* (2nd ed., pp. 285–375). Amsterdam: Academic Press.

Crepaldi, D., Rastle, K., Coltheart, M., & Nickels, L. (2010). 'Fell' primes 'fall', but does 'bell' prime 'ball'? Masked priming with irregularly-inflected primes. *Journal of Memory and Language*, *63*, 83–99.

Ford, M., Marslen-Wilson, W., & Davis, M. (2003). Morphology and frequency: Contrasting methodologies. In R. Baayen & R. Schreuder (Eds.), *Morphological structure in language processing* (pp. 89–124). Berlin: de Gruyter.

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudo-word generator. *Behavior Research Methods*, *42*(3), 627–633.

Rastle, K. (2007). Visual word recognition. In M. Gaskell (Ed.), *The Oxford handbook of psycholinguistics* (pp. 71–87). Oxford: Oxford University Press.

Schmid, H., & Laws, F. (2008). Estimation of conditional probabilities with decision trees and an application to fine-grained POS tagging. In *Proceedings of COLING-08.* Manchester.

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.