Sublexical frequency measures for orthographic and phonological units in German

Authors: M. J. Hofmann[a]*, P. Stenneken[ab], M. Conrad[a], and A. M. Jacobs[a]

[a] Department of Psychology, Freie Universität Berlin

[b] Department of Psychology, Cath. University Eichstätt-Ingolstadt

*Correspondence Author:

Markus Hofmann

Allgemeine und Neurokognitive Psychologie

Habelschwerdter Allee 45, 14195 Berlin, Germany

Phone: +49(0)30-838-55633, Fax: +4930-838-55620

 E-mail: mhof@zedat.fu-berlin.de

Running Head: German orthographic and phonological sublexical measures

Abstract:

Many recent studies have demonstrated the influence of sublexical frequency measures on language processing, or called for controlling sublexical measures when selecting stimulus material for psycholinguistic studies (Aichert, & Ziegler, 2005). The present study discusses which measures should be controlled for in what kind of study, and presents orthographic and phonological syllable, dual unit (bigram and biphoneme) and single unit (letter and phoneme) type and token frequency measures derived from the lemma and word form corpora of the CELEX lexical database (Baayen, Piepenbrock, & Gulikers, 1995). Additionally, we present the SUBLEX software as an adaptive tool for calculating sublexical frequency measures and discuss possible future applications. The measures and the software can be downloaded at www.psychonomic.org.

Keywords: sublexical units, grapheme, phoneme, bigram, biphoneme, syllable, orthographic, phonological

Recent studies demonstrate the influence of sublexical units on language processing (e.g. Nuerk, Rey, Graf, & Jacobs, 2000; Ziegler, & Goswami, 2005). Not only behavioral and neurocognitive findings in proficient adult readers, but also findings in subjects with acquired or developmental language disorders indicate the relevance of sublexical measures during language recognition and production.

However, to our knowledge, in contrast to word frequency measures (Baayen, Piepenbrock, & Gulikers, 1995; http://wortschatz.uni-leipzig.de/; Geyken, in press) sublexical unit frequency measures are not yet publicly available for the German language. For other languages, at least syllable frequency measures are available (Alameda, & Cuetos, 1995, and Davis, & Perea, in press, for Spanish; Stella, & Job, 2001, for Italian; Goslin, & Frauenfelder, 2000, New, Pallier, Brysbaert, & Ferrand, 2004, and www.lexique.org, for French; and Leung, Law, & Fung, 2004, for Chinese). Inspired by the fact that the grain size of sublexical measures is the core topic of a recent developmental theory of skilled reading and dyslexia across languages (Goswami, & Ziegler, 2006), we found it useful to calculate sublexical frequency measures with a systematic decrease in grain size. This study thus provides orthographic and phonological syllable, dual unit (bigram and biphoneme), and single unit (letter and phoneme) type and token frequency measures, derived from the lemma and word form databases of the German CELEX lexical database (Baayen et al., 1995).

By providing highly comparable measures that were calculated by the same algorithm, we hope to inspire researchers to investigate questions that are difficult to address without these measures. Moreover, we provide further independent and control variables for researchers that investigate language processing.

We start with a short overview of the fields of research in which the role of sublexical units was recently investigated, and draw particular attention to connectionist models that can account for these hypothetical levels of representation. For that purpose we outline empirical and theoretical contributions to the research fields of word recognition and naming in proficient readers, as well as of acquired and developmental language disorders. Since most of the recent studies within those fields investigate syllable frequency effects, we focus on these sublexical effects.

Carreiras, Álvarez and De Vega (1993) showed that syllable frequency plays a significant role during visual word recognition. They found that words with high frequency initial syllables take more time to be processed than words with low frequency syllables. This finding led to the hypothesis that syllables activate competing lexical candidates during lexical access. The processing delay due to syllable frequency was interpreted as interference of other lexical candidates activated by the target's syllabic units. Perea and Carreiras (1998) provided evidence that higher frequency syllabic neighbors are the source of this inhibitory syllable frequency effect. These initial findings from the Spanish language were replicated in French (Conrad, Grainger, & Jacobs, in press; Mathey, & Zagar, 2002) and German (Conrad, & Jacobs, 2004).

Whereas the effect of syllable frequency was always inhibitory in tasks requiring lexical access such as lexical decision or perceptual identification (Conrad, & Jacobs, 2004), it has been described to be either facilitative (Perea, & Carreiras, 1998) or inhibitory (Carreiras et al., 1993; Conrad, Stenneken, & Jacobs, in press) in the naming task.

Further evidence for the relevance of syllabic processing in naming and word recognition comes from eye movement measures (Carreiras, & Perea, 2004; Hutzler, Conrad, & Jacobs, 2005) and electrophysiological findings (Barber, Vergara, & Carreiras, 2004; Hutzler, Bergmann, Conrad, Kronbichler, Stenneken, & Jacobs, 2004).

The electrophysiological findings shed light on the neurocognitive processes involved in sublexical unit processing in proficient readers. It should be noted that behavioral findings are also able to contribute to the knowledge about the neuropsychology of sublexical word processing. That is, for instance, when acquired impairments of written (Stenneken, Conrad, Hutzler, Braun, & Jacobs, 2005) or spoken (Aichert, & Ziegler, 2004, Laganaro, 2005; Stenneken, Hofmann, & Jacobs, 2005; Stenneken, Bastiaanse, Huber, & Jacobs, 2005) language are compared to unimpaired functioning.

Conrad and Jacobs (2004), as well as Hutzler et al. (2004) pointed out that the syllable frequency effect provides a challenge to future computational models of word recognition, as no current model is able to account for these findings because of the lack of data on syllabic units (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Grainger, & Jacobs, 1996; Jacobs, Graf, & Kinder, 2003; Jacobs, Rey, Ziegler, & Grainger, 1998; Ziegler, Perry, & Coltheart, 2003; Zorzi, Houghton, & Butterworth, 1998; but see Ans, Carbonnel, & Valdois, 1998).

In contrast, the language production literature has provided one computational model (Levelt, Roelofs, & Meyer, 1999) that could account for syllable frequency effects (Cholin, Levelt, & Schiller, 2006). Levelt et al.'s (1999) model proposed that syllabic processing follows lexical selection that can be associated with lexical access.

Thus, it is not fully applicable to the field of word recognition in which sublexical processes also precede lexical access (Hutzler et al., 2004).

In contrast to the syllabic level of representation, smaller sized unit frequency effects have been addressed by connectionist models of word recognition and have been discussed as two of the multiple levels of representation (Grainger, & Jacobs, 1993; Grainger, & Jacobs, 1996, Massaro, & Cohen, 1994; Nuerk et al., 2000, Jacobs et al., 1998).

Much as for syllabic processing in proficient readers, there is also no computational model that could provide quantitative predictions concerning impaired syllabic processing. However, there is a pre-quantitative theory that allows for describing the proficient and impaired development of sublexical representations in different languages. Ziegler and Goswami's (2005) grain size theory emphasized the relevance of these multiple levels of sublexical unit representations for the research and treatment of dyslexia. One of the core notions of this theory is the problem of granularity. That is, the larger the sublexical units are the more of them exist. With regard to reading performance, the most economic strategy with the lowest memory effort would therefore be to link graphemes to phonemes, because for reading acquisition it is necessary to assign a phonological representation to a printed word. In the German language this is a suitable reading strategy, since graphemes usually map to only one phoneme[i] (Goswami, Ziegler, Dalton, & Schneider, 2003; Jacobs, 2002; Jacobs, & Graf, 2005; Ziegler, Perry, Jacobs, & Braun, 2001). However, in languages with more inconsistent grapheme-to-phoneme correspondences (GPC) larger units may be more suitable for reading acquisition. In some languages such as English this inconsistency consists mainly of the fact that graphemes can be spelled in multiple ways (i.e. feedforward inconsistency; Ziegler, Stone, & Jacobs, 1997). In other

languages, such as French, the main source of inconsistency consists of the fact that phonemes can be written in multiple ways (i.e. feedbackward inconsistency; Ziegler, Jacobs, & Stone, 1996).

The development of lexical and sublexical representations during language acquisition can be opposed to the most economic reading acquisition strategy, where the use of the smallest grain size appears to be most suitable. The word level representation is learned first, a syllabic representation develops usually at the age of four to five, and the representation of graphemes and phonemes develops not until reading acquisition (Ziegler, & Goswami, 2005).

The differential development of grain size representations during language and reading acquisition, as well as language specific factors that determine the most economic grain size usage strategies suggest that the question "Is there a need to control for sublexical frequencies?" (Aichert, & Ziegler, 2005) has to be answered positively. The measures of the present study could be used to build models that can make quantitative predictions concerning sublexical processes during imparired or unimpaired language processing.

Grain Sizes, Domains, Databases and Measures

The multiple grain size theory emphasizes the importance of multiple grain sizes when written words have to be mapped to phonology. The next logical step is to provide the frequencies at different grain size levels - syllables, dual units, and single units – in order to be able to address the question to what degree readers differ with respect to the reliance on different grain size units during language processing.

These three different grain size frequencies can be calculated for different domains (orthographic vs. phonological), different basic databases (word form vs. lemma), and as type and token measures. Earlier studies either were based on a subset of the frequency tables presented in the present study, or provided only incomplete information about these different possibilities to calculate frequency measures. Moreover, when predictive properties of different similar measures have to be assessed, it seems reasonable to calculate all measures in a comparable way by the same algorithm.

The present study demonstrates the diversity of ways to calculate sublexical frequency measures. However, when a researcher finally has to choose which of the proposed frequency measures to use, several issues should be considered concerning grain sizes (syllable, dual unit and single unit), processing domains (orthographic or phonological), databases (lemma or word form), and type or token measures. In the following paragraphs, we describe studies that compared the respective influences of different grain sizes on language processing. In addition, we discuss which database, domain or measure should be used for what type of study. These sections can be used as a guide when decisions for particular frequency tables have to be made.

Grain sizes: syllable, dual unit or single unit. A reliable inhibitory effect of the first syllables' frequency on lexical decisions was found reliable when bigram frequency was held constant (Conrad, Grainger & Jacobs, in press; Conrad, Carreiras, & Jacobs, in prep.). Given recent evidence that the syllable frequency effect in speech production (Cholin et al., 2006) and lexical decision (Conrad, Grainger, & Jacobs, in press) is based on the phonological syllable, biphoneme frequency might be an interesting control variable for further research. In the orthographic domain, there is

evidence for a facilitatory bigram frequency effect during lexical decision (Massaro, & Cohen, 1994), even when syllable frequency was controlled for (Conrad et al., in prep.). Moreover, Grainger and Jacobs (1993) demonstrated that letter and bigram priming effects during lexical decision are greater when units occurred at the same position within the prime and the target.

In addition to the question of the frequency of sublexical units, a controversy in the literature concerns the number of phonemes and syllables, during language production tasks (see Nickels, & Howard, 2004a, the reply of Martin, 2004, and the re-reply of Nickels, & Howard, 2004b). Nickels and Howard (2004a) obtained no syllable frequency effect in word production accuracy of aphasics that would have been independent of word imageability, word frequency, and the number of phonemes and clusters. Instead, they found evidence that "It's the number of phonemes that counts". They raised the controversial issue that phonemes are the most important units of speech production, and that effects of the phonological syllable could be attributed to confounding variables.

Aichert and Ziegler's (2004) results neither confirmed nor contradicted this interpretation, because their word repetition experiment reporting syllable frequency effects in patients with apraxia of speech did not control for phoneme frequency. However, they confirmed the prediction of Varley and Whiteside (2001) that at the phonetic encoding level (Levelt et al., 1999) motor programs are provided for high frequency syllables.

Stenneken, Hofmann, and Jacobs (2005) reported that the phonemic jargon of an aphasic patient provided a higher correlation with phoneme frequency than with syllable frequency measures. Again, these results neither contradicted nor confirmed Nickels and Howard's (2004a) hypothesis.

The grain size theory (Ziegler, & Goswami, 2005) presumably suggests that the relative influence exerted by particular grain sizes depends on individual differences. Laganaro (2005) found evidence for this. Firstly, she found that three out of seven aphasics showed an effect of syllable frequency on substitution errors. In two of them this effect was independent of phoneme frequency. Secondly, two of the aphasic subjects showed more correct responses for nonwords composed of high frequency syllables than for nonwords composed of low frequency syllables. Thirdly, she investigated the phonemic paraphasias of one aphasic subject and found that syllable frequency influenced error rates.

In accordance with the grain size theory we propose not to neglect any grain size measure, at least when assessing language disorders. When word recognition studies are conducted, at least syllable frequency and bigram frequency should be controlled for. An independent effect of smaller sublexical measures should be evaluated to test the predictions of the grain size theory. This can be done during stimulus generation by controlling or manipulating variables, or by applying multiple regression methods in a post-hoc fashion.

Processing domains: orthography or phonology. When choosing between the orthographic and phonological domain one could suppose that written language performance can be assessed best by referring to orthographic frequency measures, and spoken language performance by phonological measures. However, particularly with regard to reading, this might be the most interesting and most controversial issue. Whereas Seidenberg (1985) claimed that phonology is not necessary for reading, Van Orden's (1987) article "a rows is a rose" presented strong arguments in favor of the

notion that phonological representations are automatically and always activated during silent reading.

Today, there seems to be a broad agreement that multiple codes are activated during reading, in particular phonological codes (Ans et al., 1998; Jacobs et al., 1998; Yates, 2005; Ziegler, Van Orden, & Jacobs, 1997), at different grain size levels (Ziegler, & Goswami, 2005; Goswami, & Ziegler, 2006). Conrad, Grainger and Jacobs (in press) suggested that, during word recognition, it is the phonological syllable, not the orthographic syllable that drives the syllable frequency effect. This issue was investigated in a deep orthography with rather inconsistent GPC (Liberman, Liberman, Mattingly, & Shankweiler, 1980), because in shallow orthographies phonological syllable frequency is confounded with orthographic syllable frequency. In this context the question arises whether the primacy of the phonological syllable can be generalized to shallow orthographies like German, too. This question can be addressed by using regression methods in order to find out which type of syllable frequency is most predictive. Experiments using an orthogonal design, and thus manipulating orthographic and phonological syllable frequency independently, can hardly be realized in a shallow orthography.

Today, it is well accepted that a proficient reader in a language having a deep orthography can hardly avoid phonological processing when exposed to letter strings (e.g. Sumiya, & Healy, 2004). When investigating spoken language, the question arises whether highly overlearned orthographic representations of a letter string are also activated (Ziegler, & Ferrand, 1998; Ziegler, Ferrand, & Montand, 2004). If one is not interested in addressing this particular question, we suggest that the phonological domain's frequencies are used when investigating spoken language. When dealing with questions concerning reading, this choice is much more difficult.

However, the aforementioned findings suggest that using the frequencies of phonological units are as plausible as using the frequencies of orthographic units when conducting word recognition experiments

Databases: lemma or word form. CELEX (Baayen et al., 1995) provides a lemma and a word form database. The lemma database provides words in its basic form, i.e. nouns are presented in nominative singulars and verbs are presented in infinitives. In contrast, the inflected forms are provided in the word form database. Most psycholinguistic studies use the lemma database. Duyck, Desmet, Verbeke, and Brysbaert (2004) provided "WordGen", a stimulus selection tool for psycholinguistic research. The authors argued (Duyck et al., 2004. pp. 490) that they used the lemma database of CELEX, because extensive manual coding and disambiguation made the lemma database more transparent with respect to its records than the word form database. Moreover, they argued that word forms partly activate its corresponding lemma entry in the mental lexicon (Baayen, Dijkstra, & Schreuder, 1997; New, Brysbaert, Segui, Ferrand, & Rastle, 2004).

On the one hand, we agree with these arguments, in particular because Levelt et al.'s (1999) influential model proposed the fast and automatic activation of lemmas during word form processing. On the other hand, we suggest that lemma measures systematically over- or underestimate the frequency of sublexical units that occur in inflective morphemes, an issue that will be demonstrated on the basis of the results of this study. Using word form measures not only allows for evaluating language in its natural form, but it is of particular interest when, for example, sentence processing tasks are used. Thus, the choice for a certain database should be based on the task and the theoretical assumptions of a particular study.

Measures: type or token. The type measure indicates the number of words that contain the specific grain size. For example, the type frequency of the bigram "ba" denotes the number of words that contain this bigram. The token frequency, in contrast, denotes the summed frequencies of the words that contain "ba". Conrad, Carreiras and Jacobs (subm.) showed that it was the token measure of syllable frequency that appears to be responsible for the inhibitory effect of syllable frequency in lexical decision (see above). However, the authors argued that the type measure of syllable frequency led to faster response latencies especially when the number of higher frequency syllabic neighbors was controlled for.

Novick and Sherman (2004) provided two reasons for using type measures. They argued that token frequency is confounded to a large degree with word frequency, and found that type bigram frequency was a better predictor for performance in anagram resolution. However, Bailey and Hahn (2001) found that wordlikeness judgments are a function of the token frequency of lexical neighbors.

It should be noted that there is a controversial debate about the general impact of type and token measures in the current literature. Many of the contributions to this debate describe sublexical, but neither syllabic, nor dual unit or single unit influences on language processing. De Jong, Schreuder, and Baayen (2000) found evidence that it was the type frequency of a word's root morpheme that influences response latencies in lexical decision in Dutch. Eddington (2004) found that type frequency is a better predictor than token frequency while simulating correct outcomes of Spanish stress assignment and English past tense formation. When participants had to produce a past tense ending for pseudo-verbs and verbs in Dutch they completed the words

with endings of a higher type frequency (Ernestus, & Baayen, 2003). In contrast, there was an effect of token frequency in the same paradigm (Ernestus, & Baayen, 2001).

To resolve the whole controversy, Clahsen (1999) proposed a dual route system that explains type-based analogical effects by a symbolic rule application mechanism, and token-based effects by an associative memory store. Others question the necessity of separate type- and token-sensitive mechanisms by use of connectionist models showing that the differential effects can be reduced to a single token-based mechanism (Moscoso del Prado Martín, Ernestus, & Baayen, 2004; Moscoso del Prado Martín, Kostic, & Baayen, in press). The decision for one of the measures should be based on previous research working with comparable paradigms. Useful contributions to this controversy would be to conduct a regression analysis with type and token measures as predictors, to find out which measure is most predictive, or, to manipulate type and token measures independently.

In any case, on the basis of empirical studies that compared different grain size units systematically the choice for particular frequency measures should be made. The measures of the present study offer the possibility to un-confound a large amount of variables that potentially pose a problem in interpreting results of recent studies. For example, experiments can be designed that manipulate phoneme frequency while keeping syllable frequency constant. It might help to systematically manipulate the (and only the) variables of interest. Even when investigating whole word effects, for example the emotional valence of words (e.g. Kuchinke, Jacobs, Grubich, Võ, Conrad, & Herrmann, 2005), the sublexical measures of the present study can be used to rule out the possibility that these effects might be due to the confound between sublexical measures and emotional valence.

When a researcher has to choose which of the frequency measures to use, in accordance with the grain size theory (Ziegler, & Goswami, 2005) we would suggest neither to neglect the syllabic, nor the dual unit nor the single unit grain size level. The phonological domain's frequency measures can be used, not only during the assessment of spoken language, but also while assessing written language, as suggested by the multiple code activation hypothesis (Jacobs et al., 1998). Furthermore, we suggest using word form measures in particular when assessing language as it occurs in its natural inflected form (e.g. in sentences or connected speech). Levelt et al.'s (1999) hypothesis of the automatic activation of lemma entries during word form processing also suggests using frequencies of the lemma database. One reason (see Duyck et al., 2004) to use lemma measures in particular when assessing non-inflected language may be the extensive manual coding and disambiguation within the lemma database of the CELEX lexical database (Baayen et al., 1995). When deciding whether to use either the type or the token measures, the decision should be based on prior research working with the same experimental paradigms. A better solution might be to contribute to the controversy of type vs. token measures by taking into account both of them. This could be helpful, as long as the reduction to a token based mechanism (Moscoso del Prado Martín et al., 2004, in press) has not been broadly accepted.

Method

All measures were calculated using shell scripts, PERL scripts, and the free UNIX programs join, sort and wc. Thus all software used for the present study ran under a free licence. A Macintosh G4 computer was used running a free BSD under Mac Os X 10.3.9, as the native operating system of the SUBLEX-software[ii]. However, SUBLEX should run on every UNIX or LINUX shell running with an ISO Latin 9 character set.

Each step of calculation can be adapted flexibly, for example to calculate case-sensitive measures (see README.txt). At this point we will give an overview about all processing steps and provide the results when the program is executed without modifications. The program and the resulting frequency measures can be downloaded at www.psychonomic.org.

The sublexical measures were derived from the German orthographic lemmas, the German phonological lemmas, the German orthographic word forms, and the German phonological word forms of the CELEX lexical database (Baayen et al., 1995).

Words with acute accents (/#/) were identified as foreign words from the orthographic lemma and word form databases, and excluded from analysis. The phonological transcription of the CELEX[iii] was used to exclude words that contained a phoneme occurring only in other languages than German. Words that contained a /~/, an /A/, a /Z/, an /O:/, an /3:/, a /w/, or an /V/ were excluded from analyses. All words that contained a shortly pronounced /e/ or /&/ were excluded from analyses. Additionally, the orthographic and phonological syllable number of each entry was compared. In order to exclude foreign words and errors of the phonological transcription, entries with different orthographic and phonological syllable numbers

were excluded from analysis. 51,207 words remained in the lemma database for analysis.

The 363,013 entries of the adjusted word form database consisted of words and phrases (e.g. "bestelltest ab"). Phrases in which the number of words differed in the orthographic and phonological notation were excluded from analysis. The words of a phrase were processed as separate words, with the respective word frequency of the whole phrase. 44,033 phrases consisted of two words and 315 phrases consisted of 3 words. After foreign words have been excluded from analysis, 407,676 words remained in the adjusted word form database.

For the calculation of all phonological sublexical measures long vowels (/a:/, /E:/, /e:/, /i:/, /o:/, /u:/, /y:/, and /&:/) were treated differently from short vowels (/a/, /E/, /e/, /i/, /o/, /u/, and /y/). When one wants to neglect this distinction, long and short vowel frequencies can be summed post-hoc. To calculate the phonological syllable frequencies, ambisyllabic consonants were attributed to both syllables.

All uppercase letters were converted to lowercase, to obtain case insensitive frequencies. The resulting type frequency measures indicate the number of times a sublexical unit occurs in the respective CELEX database. The token measures refer to the sum of the CELEX's Mannheim frequency of the lexical entries that contained this particular unit. Token frequency measures are given in occurrence per 6 million.

Results

The complete syllable, dual unit (bigram and biphoneme) and single unit (letter and phoneme) type and token frequency measures that were calculated for different domains (orthographic vs. phonological) and different basic databases (word

form vs. lemma) are available at www.psychonomic.org (see README.txt for the nomenclature of the files). Here, we will illustrate the findings by providing the most frequent sublexical units. For syllable and dual unit frequencies we will additionally provide the number and one example of the most rare sublexical units, respectively. For single unit frequencies, we describe the rarest letters and phonemes.

Syllable frequencies of the lemma database. 6,023 different orthographic and 5,679 different phonological syllables were extracted from the 163,099 orthographic and phonological syllables of the lemma database.

The orthographic and phonological syllable with the highest *type* frequency was "ge" and /g@/. It occurred in 3,076 orthographic and 2,561 phonological words. The derivative affixes "ver" (/fEr/) and "be" (/b@/) were the only other syllables that occurred in more than 2,000 orthographic and phonological words. There were 1,529 orthographic and 1,315 phonological syllables that occurred in only one word (e.g. the free morpheme "auch" or /aux/ was never a syllable of another word than itself).

The orthographic and phonological syllable with the highest *token* frequency was "der" (/de:r/). The summed frequency of all words that contained this syllable was 703,722 orthographically and 660,055 phonologically. The only syllables with an orthographic and phonological token frequency larger than 150,000 was "und", while "ge" exceeded this criterion only orthographically. There were 843 orthographic and 724 phonological syllables that occurred only in words with a CELEX word frequency of zero (e.g. "sext" and /zEkst/ occurred only in words like Sextakkord, /zEkstakOrt/).

Syllable frequencies of the word form database. 11,731 orthographic syllables and 10,772 different phonological syllables were derived from the 1,285,294 syllables of the word form database.

Again, "ge" and /g@/ were the syllables with the highest *type* frequency (orthographic: 35,743, phonological: 30,585). The only other phonological syllables that occurred in more than 20,000 words were /t@n/ and /t@/. Orthographically, "te" reached this criterion and "ten" marginally missed it with a frequency of 19,691. There were 2,231 orthographic and 1,837 phonological syllables that occurred only in one word (e.g. /o:l/ from /Spani:o:l/, Spaniol; "auch" see above).

Again, "der" had the highest orthographic *token* frequency and /de:r/ had the second highest (orthographic: 269,011, phonological: 219,912). The word with the highest phonological token frequency and the second highest orthographic token frequency was /di:/ ("die") with a summed frequency of 240,694, orthographically, and 249,273 phonologically. There were 4,470 orthographic and 3,841 phonological syllables that occurred only in words with a frequency of zero (e.g. /E:rst/ from /fami:li:E:rst/, familiärst, or "brückst" from "überbrückst").

Dual unit frequencies of the lemma database. 710 different bigrams and 979 different biphonemes were derived from the 453,770 bigrams and 411,358 biphonemes of the lemma database.

The bigram with the highest *type* frequency was "er" (17,315), followed by "en" and "ch" as the only bigrams that occurred in more than 10,000 words. There were only 27 bigrams that occurred in only one word (e.g. "gc" from "Spängchen"). The biphoneme that occurred in the largest number of words (13,756) was /@n/, followed by /@r/ as the only other biphoneme that occurred in more than 9,000

words. There were 47 biphonemes that occurred in only one word (e.g. /zv/ from /SErzvaiz@/, scherzweise).

The bigram with the highest *token* frequency was "er", too (summed frequency of 1,487,559). "en" was the only other bigram with a token frequency higher than 1,000,000. There were 20 bigrams that occurred only in words with a frequency of zero (e.g. "vl" from "Frevler"). The biphoneme with the highest token frequency (896,914) was /@n/. The only other biphoneme that had a higher token frequency than 80,000 was /e:r/. There were 29 biphonemes that only occurred in words with a frequency of zero (e.g. /E:h/, /zE:hait/, Zäheit).

Dual unit frequencies of the word form database. 721 different bigrams and 993 different biphonemes were derived from the 3,579,388 bigrams and 3,273,332 biphonemes of the word form database.

Again, the bigram with the highest *type* frequency (160,582) was „er". The bigrams „ch", „st", "en", and „te" occurred in more than 100,000 words. Three bigrams occurred only in one word (e.g. „cc" from „staccato"). The biphoneme that occurred in the largest number of words (121,822) was /t@/, followed by /@n/ as the only other biphoneme occurring in more than 100,000 words. 13 biphonemes occurred only in one word (e.g. /io:/, see above).

The bigram with the highest *token* frequency (1,048,911) was „en". The only other bigram that had a higher token frequency than 1,000,000 was „er". 26 bigrams occurred only in words with a frequency of zero (e.g. „cc", see above). The biphoneme with the highest token frequency (841,141) was /@n/. /ai/ was the only other biphoneme exceeding the 500,000 token frequency threshold. 35 biphonemes occurred only in words with a frequency of zero (e.g. /E:h/, see above).

Single unit frequencies of the lemma database. 30 different letters and 38 different phonemes were derived from the 505,028 letters and 462,613 phonemes of the lemma database

The letter with the highest *type* frequency was "e" (69,860). The only other letters that occurred in more words than 40,000 were "n" and "r". The letters "q", "x", "j" and "y" occurred in less than 1,000 words. The phoneme that occurred in the largest amount of words (40,725) was /t/, followed by /r/, /n/, /@/, and /a/ which exceeded the 30,000 words threshold. The only phonemes that occurred in less than 1,000 words were /j/ and /Q/.

The letter with the highest *token* frequency was "e" (4,411,788). The only letters that exceeded the token frequency threshold of 2,000,000 were "n" and r". The letters "q", "x", and "y" had a token frequency below 10,000. The phoneme with the highest token frequency was /n/ (2,545,874). The only other phoneme that exceeded the 2,000,000 threshold was /r/. The phonemes with the lowest token frequency were /&:/ and /Q/, that had a token frequency below 50,000.

Single unit frequencies of the word form database. Again, 30 different letters and 38 different phonemes were derived from the 3,987,164 letters and 3,681,103 phonemes of the word form database.

The letter with the highest *type* frequency was "e" (663,642). The letters "t", "s", and "r" occurred in more words than 300,000. The only letters that occurred in less words than 10,000 were "q", "j", "x" and "y". The phoneme that occurred in the largest amount (431,585) of words was /@/. The only other phoneme occurring in

more words than 400,000 was /t/. The phonemes /j/, /Q/ and /&:/ occurred in less than 10,000 words.

Again, the letter with the highest *token* frequency (4,595,079) was "e". Letters that had a higher token frequency than 2,000,000 were "n", "i", and "r". The only letters that had a token frequency lower than 10,000 were "q", "x", and "y". The phoneme with the highest token frequency was /n/ (2,504,247). The only other phonemes with a token frequency higher than 2,000,000 were /@/ and /t/. The only phonemes with a token frequency lower than 100,000 were /Q/, /&:/, /Y/, /E:/, /y/, and /j/.

In German, most phonemes correspond to one letter. However, there are a few two-letter units (e.g. ch, ck). These frequency counts can be derived from the respective frequency lists at www.psychonomic.org. In order to allow the assessment of the frequencies of all German graphemes, we also provide the frequency of the only three-letter grapheme here: "sch". The *type* frequency of "sch" was 7,082 in the lemma corpus, and 54,270 in the word form corpus. The *token* frequency was 228,422 in the lemma corpus and 228,414 in the word form corpus.

Discussion

Whereas earlier studies assessed sublexical frequency effects based on lemma corpora (e.g. Conrad, & Jacobs, 2004), or did not specify from which corpus the measures were derived, the present study provides also sublexical frequency measures derived from the German word form corpora (Baayen et al., 1995). Hence, sentence-level studies can be conducted avoiding the systematic over- or underestimation of

syllable frequencies that determine the inflection of a word that would result from using the lemma database. For example, the word form "wird" contributes to the lemma frequency of the word "werden", and thus the frequency measures of syllables "wer" and "den" are systematically overestimated. On the other hand, underestimations can occur, for example, in syllables that correspond to inflective morphemes. Thus, the syllables "ten" (/t@n/) and „te" (/t@/) that correspond to the German past tense inflective morphemes, are much more frequent in the word form than in the lemma database.

All syllabic level analyses provided more orthographic than phonological syllables. This pattern of results shows that the German language is more feedbackward than feedforward inconsistent (Stone, Vanhoy, and Van Orden, 1997; Ziegler et al., 1996; Ziegler, Stone, & Jacobs, 1997). Orthographic syllables necessarily have to be spelt in different ways to generate this number relation. One source of this inconsistency in German is the fact that there is often no orthographic difference between vowels that are pronounced long or short, e.g. the orthographic syllable "ol" corresponds to one phonological syllable when it is pronounced long (/o:l/), and to 17 words when it is pronounced short (/ol/). The calculation of such inconsistencies is one example of additional measures that can be derived from the CELEX lexical database by making small modifications to the SUBLEX software.

When it comes to smaller sublexical units, not only the summed positional bigram measures can be derived from the data of the present study (see Duyck et al., 2004), but also the mean bigram frequency of a word which is not confounded with word length. Additionally, the present study provides the first online database of biphoneme measures.

Once phoneme and syllable frequency measures are available in this way, every study that investigates the processing of word stimuli can, in principle, contribute to Nickels and Howard's (2004a) controversy (see above) that raised the question whether syllable frequency has an influence that is independent of phoneme frequency. This can be done either by controlling for either of both variables, or by evaluating the independence of effects by applying multiple regression methods.

All measures are now provided by one study, and were calculated by the same algorithm. Thus, it is now possible to compare the relative influences of each measure in different tasks. It also becomes possible to evaluate which group of subjects is sensitive to what degree to which sublexical measure in which task. The present study has provided the basic data to meet Aichert and Ziegler's (2005) call for controlling sublexical measures. Systematic comparisons between syllable, dual unit and single unit measures, between the orthographic and the phonological domain, between type and token measures, as well as between measures derived from the lemma and word form database are now possible.

It is well known that the larger the grain size, the more units exist (Ziegler, & Goswami, 2005). Now, concrete numbers are available for German. The CELEX lexical database (Baayen et al., 1995) consists of 10,722 syllables, 979 biphonemes and 38 phonemes, derived from the German words of the CELEX word form corpus. According to CELEX, German written texts contain 11,731 syllables, 710 bigrams, and 30 letters.

It should be noted that the present study neglected positional frequency measures (in contrast to Massaro's & Cohen's, 1994, approach to bigram frequency, for instance), and concentrated on non-positional measures (as e.g. Duyck et al., 2004). The grain size units were counted irrespective of the position in a word. The

question which of both measures reflects the processing of a stimulus best has not yet been answered to our knowledge. Position specificity is a matter of debate in the current literature that has more than these two solutions (see Dehaene, Cohen, Sigman, & Vinckier, 2005; Goswami, & Ziegler, 2006; Grainger, & Whitney, 2004). For example, relative positions within a word might be another suitable concept (Peressotti, & Grainger, 1999). Thus, we decided to neglect position specificity for the present purposes.

        To find out how sublexical frequency measures can be applied to the diagnosis of language skills, Seidenberg's (1987) principle of orthographic redundancy can be used in a developmental perspective of reading or language abilities in general (Seidenberg, & Mc Clelland, 1989). Not all orthographic patterns are equally frequent. Thus, orthographic patterns that occur very rarely are less likely to be recognized than high frequency patterns. We propose that the present studies' frequency measures can be used to determine the relative reliance on particular grain sizes during reading or speaking of an individual. By manipulating each grain size and holding the respective other grain sizes constant, a certain frequency for each grain size and participant can be obtained. Hypothetically, units above these diagnostically relevant frequencies are processed correctly, in contrast to units below that frequency.

        By knowing the relative strengths of an impaired reader during the processing of a particular grain size, compensational strategies can be taught to generalize from the relatively impaired grain sizes to other grain sizes, if proficient reading is correctly characterized by the activation of multiple grain sizes (Ziegler, & Goswami, 2005).

        Another therapeutic approach deals with the fact that small units are learned by finding the differences between large units (Ziegler, & Goswami, 2005). For instance, by naming the common phonemes in the words /pa:t@/ and /kOst/ a reader

can gain a cognitive representation of the phoneme /t/. On the basis of the present analysis therapeutic strategies should initially use high frequency phonemes in unskilled readers that can be learned easier than lower frequent phonemes.

The calculation algorithms now being available could be used to calculate these measures for other languages provided by the CELEX lexical database (English and Dutch). Such follow-up analyses could easily be performed by a slightly modified SUBLEX software. Since the grain size theory can also contribute to a cross-linguistic perspective (Ziegler, & Goswami, 2005), such follow-up studies would allow for comparing the relative influence of different grain sizes across languages. Ziegler and Goswami (2005) already predicted that in languages with more inconsistent GPC (e.g. English) larger grain size units might be more suitable than in languages with more consistent GPC. Such hypotheses can be tested by use of the materials provided by such follow-up studies.

We hope that the SUBLEX software will also be applied to newer corpora of the German language, such as the Web-CELEX (see http://www.mpi.nl/world/celex/), the DWDS-corpus (Geyken, in press), or the German Wortschatz-Project (http://wortschatz.uni-leipzig.de/).

Acknowledgements

Authors Note

None of the authors has any financial interest concerning any materials presented in the paper.

References

Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, **88**, 148-159.

Aichert, I., & Ziegler, W. (2005). Is there a need to control for sublexical frequencies? *Brain and Language*, **95**, 170-171.

Alameda, J. R., & Cuetos, F. (1995). *Diccionario de frecuencia de las unidades lingüísticas del castellano, Vols. I & II*. Oviedo: Universidad de Oviedo.

Ans, B., Carbonnel, S., & Valdois, S. (1998). A Connectionist Multiple-Trace Memory Model for Polysyllabic Word Reading. *Psychological Review*, **105**, 678-723.

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

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (CD-ROM):* Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Bailey, T. M., & Hahn, U. (2001). Determinants of Wordlikeliness: Phonotactics or Lexical Neighborhoods? *Journal of Memory and Language*, **44**, 568-591.

Barber, H., Vergara, M., & Carreiras, M. (2004). Syllable-frequency effects in visual word recognition: Evidence from ERPs. *NeuroReport*, **15**, 545-548.

Carreiras, M., Álvarez, C. J., & De Vega, M. (1993). Syllable frequency and visual word recognition in Spanish. *Journal of Memory and Language*, **32**, 766-780.

Carreiras, M., & Perea, M. (2004). Effects of syllable neighbourhood frequency in visual word recognition and reading: Cross-task comparisons. In L. Ferrand & J. Grainger (Eds.), *Psycholinguistique Cognitive*. Bruxelles: De Broeck Université.

Cholin, J., Levelt, W. J. M., & Schiller, N. O. (2006). Effects of syllable frequency in

speech production. *Cognition*, **99**, 205-235.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavioral and brain sciences*, **22**, 991-1060.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, **108**, 204-256.

Conrad, M., Carreiras, M., & Jacobs, A. M. (in preparation). Syllable frequency and orthographic redundancy evidence for two different processing mechanisms in visual word recognition.

Conrad, M., Carreiras, M., & Jacobs, A. M. (submitted). Contrasting effects of token and type syllable frequency in lexical decision.

Conrad, M., Grainger, J., & Jacobs, A. (in press). Phonology as the source of syllable frequency effects in visual word recognition: Evidence from French. *Memory and Cognition*.

Conrad, M., & Jacobs, A. M. (2004). Replicating syllable frequency effects in Spanish in German: One more challenge to computational models of visual word recognition. *Language and Cognitive Processes*, **19**, 369-390.

Conrad, M., Stenneken, P., & Jacobs, A. M. (in press). Associated or dissociated effects of syllable-frequency in lexical decision an naming. *Psychonomic Bulletin and Review*.

Davis, C., & Perea, M. (in press). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, **9**, 335-341.

De Jong, N. H., Schreuder, R., & Baayen, R. H. (2000). The morphological family size

effect and morphology. *Language and Cognitive Processes*, **15**, 329-365.

Duyck, W., Desmet, T., Verbeke, L. P. C., & Brysbaert, M. (2004). WordGen: A tool for word selection and nonword generation in Dutch, English, German, and French. *Behavior Research Methods, Instruments, and Computers*, **36**, 488-499.

Eddington, D. (2004). Issues in modeling language processing analogically. *Lingua*, **114**, 849-871.

Ernestus, M., & Baayen, R. H. (2001). Choosing between the Dutch past-tense suffixes -te and -de. In T. van der Wouden (Eds.), *Linguistics in the Netherlands 2001*. (pp. 81-93). Amsterdam: Benjamins.

Ernestus, M., & Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, **79**, 5-38.

Geyken, A. (in press). The DWDS-corpus: A reference corpus for the German language of the 20th century. In C. Fellbaum (Eds.), *Collocations and Idioms: Linguistic, Lexicographic, and Computational Aspects*. London: Continuum Press.

Goslin, J., & Frauenfelder, U. H. (2000). A comparison of theoretical and human syllabification. *Language and Speech*, **44**, 409_436.

Goswami, U., & Ziegler, J. (2006). A developmental perspective on the neural code for written words. *Trends in Cognitive Sciences*, **10**, 142-143.

Goswami, U., Ziegler, J. C., Dalton, L., & Schneider, W. (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics*, **24**, 235-247.

Grainger, J., & Jacobs, A. M. (1993). Masked partial-word priming in visual word recognition: effects of positional letter frequency. *Journal of Experimental Psychology: Human Perception and Performance*, **19**, 951-964.

Grainger, J., & Jacobs, A. M. (1996). Orthographic Processing in Visual Word Recognition:

A Multiple Read-Out Model. *Psychological Review*, **103**, 518-565.

Grainger, J., & Whitney, C. (2004). Does the huamn mnid raed wrods as a wlohe? *Trends in Cognitive Sciences*, **8**, 58-59.

Hutzler, F., Bergmann, J., Conrad, M., Kronbichler, M., Stenneken, P., & Jacobs, A. M. (2004). Inhibitory effects of first syllable-frequency in lexical decision: An event-related potential study. *Neuroscience Letters*, **372**, 179-184.

Hutzler, F., Conrad, M., & Jacobs, A. M. (2005). Effects of syllable-frequency in lexical decision and naming: An eye-movement study. *Brain and Language*, **92**, 138-152.

Jacobs, A. M., Graf, R., & Kinder, A. (2003). Receiver Operating Characteristics in the Lexical Decision Task: Evidence for a Simple Signal-Detection Process Simulated by the Multiple Read-Out Model. *Journal of Experimental Psychology: Learning Memory and Cognition*, **29**, 481-488.

Jacobs, M., & Graf, R. (2005). Wordform memory as intuitive statistics in languages with different consistency [Wortformgedaechtnis als intuitive Statistik in Sprachen mit unterschiedlicher Konsistenz]. *Zeitschrift für Psychologie*, **213**, 133-141.

Jacobs, A. M., Rey, A., Ziegler, J. C., & Grainger, J. (1998). MROM-P: An interactive activation, multiple read-out model of orthographic and phonological processes in visual word recognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition.* (pp. 147-187). Mahwah, NJ: Erlbaum.

Jacobs, A. M. (2002). The cognitive psychology of literacy. In N. J. Smelser & P. B. Baltes (Eds.), *International Encyclopedia of the Social and Behavioral Sciences.* (pp. 8971-8975). Amsterdam: Elsevier.

Kuchinke, L., Jacobs, A. M., Grubich, C., Võ, M. L.-H., Conrad, M., & Herrmann, M. (2005). Incidental effects of emotional valence in single word processing: An fMRI study. *NeuroImage*, **28**, 1022-1032.

Laganaro, M. (2005). Syllable frequency effect in speech production: Evidence from aphasia. *Journal of Neurolinguistics*, **18**, 221-235.

Leung, M.-T., Law, S.-P., & Fung, S.-Y. (2004). Type and token frequencies of phonological units in Hong Kong Cantonese. *Behavior Research Methods, Instruments, and Computers*, **36**, 500-505.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, **22**, 1-75.

Liberman, I., Liberman, A. M., Mattingly, I., & Shankweiler, D. (1980). *Orthography and the beginning reader*. Baltimore, OH: University Park Press.

Martin, N. (2004). Comments on Nickels and Howard (2004) "Dissociating effects of number of phonemes, number of syllables, and syllable complexity on word production in aphasia: It's the number of phonemes that counts". *Cognitive Neuropsychology*, **21**, 528-530.

Massaro, D. W., & Cohen, M. M. (1994). Visual, orthographic, phonological, and lexical influences in reading. *Journal of Experimental Psychology: Human Perception and Performance*, **20**, 1107-1128.

Mathey, S., & Zagar, D. (2002). Lexical similarity in visual word recognition: The effect of syllabic neighborhood in French. *Current Psychology Letters: Behavior, Brain & Cognition*, **8**, 107-121.

Moscoso del Prado Martín, F., Ernestus, M., & Baayen, R. H. (2004). Do type and token reflect different mechanisms? Connectionist modeling of Dutch past-tense formation and final devoicing. *Brain and Language*, **90**, 287-298.

Moscoso del Prado Martín, F., Kostic, A., & Baayen, R. H. (in press). Putting the bits together: An information theoretical perspective on morphological processing. *Cognition*.

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

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments, and Computers*, **36**, 516-524.

Nickels, L. A., & Howard, D. (2004a). Dissociating effects of number of phonemes, number of syllables, and syllabic complexity on word production in aphasia: It's the number of phonemes that counts. *Cognitive Neuropsychology*, **21**, 57-78.

Nickels, L. A., & Howard, D. (2004b). Correct responses, error analyses, and theories of word production: A response to Martin. *Cognitive Neuropsychology*, **21**, 531-536.

Novick, L. R., & Sherman, S. J. (2004). Type-based bigram frequencies for five-letter words. *Behavior Research Methods, Instruments, and Computers*, **36**, 397-401.

Nuerk, H.-C., Rey, A., Graf, R., & Jacobs, A. M. (2000). Phonographic sublexical units in visual word recognition. *Current Psychology Letters: Behavior, Brain, & Cognition*, **2**, 27-38.

Perea, M., & Carreiras, M. (1998). Effects of Syllable Frequency and Syllable Neighborhood Frequency in Visual Word Recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **24**, 134-144.

Peressotti, F., & Grainger, J. (1999). The role of letter identity and letter position in orthographic priming. *Perception and Psychophysics*, **61**, 691-706.

Seidenberg, M. (1985). The time course of phonological code activation in two writing systems. *Cognition*, **19**, 1-30.

Seidenberg, M. S. (1987). Sublexical structures in visual word recognition: Access units or orthographic redundancy? In M. Coltheart (Eds.), *Attention and performance XII*. (pp. 245-263). Hillsdale, KJ: Erlbaum.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word

recognition and naming. *Psychological Review*, **96**, 523-568.

Stella, V., & Job, R. (2001). Le sillabe PD/DPSS. Una base di dati sulla frequenza sillabica

dell'italiano scritto. *Giornale Italiano di Psicologia*, **28**, 633_639.

Stenneken, P., Conrad, M., Hutzler, F., Braun, M., & Jacobs, A. M. (2005). Frequency

effects with visual words and syllables in a dyslexic reader. *Behavioral Neurology*,

16, 1-15.

Stenneken, P., Hofmann, M., & Jacobs, A. M. (2005). Patterns of phoneme and syllable

frequency in jargon aphasia. *Brain and Language*, **95**, 221-222.

Stenneken, P., Bastiaanse, R., Huber, W. & Jacobs, A.M. (2005). Syllable structure and

sonority in language inventory and aphasic neologisms. *Brain and Language,* **95**, 280-

292.

Stone, G. O., Vanhoy, M., & Van Orden, G. C. (1997). Perception is a two-way street:

Feedforward and feedback phonology in visual word recognition. *Journal of Memory

and Language*, **36**, 337-359.

Sumiya, H., & Healy, A. F. (2004). Phonology in the bilingual Stroop effect. *Memory and

Cognition*, **32**, 752-758.

Van Orden, G. C. (1987). A ROWS is a ROSE: spelling, sound, and reading. *Memory &

Cognition*, **15**, 181-198.

Varley, R., & Whiteside, S. P. (2001). What is the underlying impairment in acquired

apraxia of speech? *Aphasiology*, **15**, 39-49.

Yates, M. (2005). Phonological neighbors speed visual word processing: Evidence from

multiple tasks. *Journal of Experimental Psychology: Learning Memory and

Cognition*, **31**, 1385-1397.

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of

orthographic consistency on different auditory word recognition tasks. *Memory and Cognition*, **32**, 732-741.

Ziegler, J. C., & Ferrand, L. (1998). Orthography shapes the perception of speech: The consistency effect in auditory word recognition. *Psychonomic Bulletin and Review*, **5**, 683-689.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin*, **131**, 3-29.

Ziegler, J. C., Jacobs, A. M., & Stone, G. O. (1996). Statistical analysis of the bidirectional inconsistency of spelling and sound in French. *Behavior Research Methods, Instruments, and Computers*, **28**, 504-515.

Ziegler, J. C., Perry, C., & Coltheart, M. (2003). Speed of lexical and nonlexical processing in French: The case of the regularity effect. *Psychonomic Bulletin and Review*, **10**, 947-953.

Ziegler, J. C., Perry, C., Jacobs, A. M., & Braun, M. (2001). Identical words are read differently in different languages. *Psychological Science*, **12**, 379-384.

Ziegler, J. C., Stone, G. O., & Jacobs, A. M. (1997). What is the pronunciation for -ough and the spelling for /u/? A database for computing feedforward and feedback consistency in English. *Behavior Research Methods, Instruments, and Computers*, **29**, 600-618.

Ziegler, J. C., Van Orden, G. C., & Jacobs, A. M. (1997). Phonology Can Help or Hurt the Perception of Print. *Journal of Experimental Psychology: Human Perception and Performance*, **23**, 845-860.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two Routes or One in Reading Aloud? A Connectionist Dual-Process Model. *Journal of Experimental Psychology: Human*

Perception and Performance, **24**, 1131-1161.

---

[i] There are exceptions such as terminal devoicing and the fact that there is often no orthographic differentiation between short or long spoken vowels.

[ii] The program is distributed under a free GNU-licence.

[iii] The syllabified phonological headwords in the CELEX charset.