The grapho-phonological system of written French: Statistical analysis and empirical validation

Marielle Lange
Laboratory of Experimental Psychology,
Université Libre de Bruxelles
Av. F.D. Roosevelt, 50
Bruxelles, Belgium, B 1050 Bruxelles
mlange@ulb.ac.be

Alain Content
Laboratory of Experimental Psychology,
Université Libre de Bruxelles
Av. F.D. Roosevelt, 50
Bruxelles, Belgium, B 1050 Bruxelles
acontent@ulb.ac.be

Abstract

The processes through which readers evoke mental representations of phonological forms from print constitute a hotly debated and controversial issue in current psycholinguistics. In this paper we present a computational analysis of the grapho-phonological system of written French, and an empirical validation of some of the obtained descriptive statistics. The results provide direct evidence demonstrating that both grapheme frequency and grapheme entropy influence performance on pseudoword naming. We discuss the implications of those findings for current models of phonological coding in visual word recognition.

Introduction

One central characteristic of alphabetic writing systems is the existence of a direct mapping between letters or letter groups and phonemes. In most languages, although to a varying extent, the mapping from print to sound can be characterized as quasi-systematic (Plaut, McClelland, Seidenberg, & Patterson, 1996; Chater & Christiansen, 1998). Thus, descriptively, in addition to a large body of regularities (e.g. the grapheme CH in French regularly maps onto /ʃ/), one generally observes isolated deviations (e.g. CH in CHAOS maps onto /k/) as well as ambiguities. In some cases but not always, these difficulties can be alleviated by considering higher order regularities such as local orthographic environment (e.g., C maps onto /k/ or /s/ as a function of the following letter), phonotactic and phonological constraints as well as morphological properties (Cf. PH in PHASE vs. SHEPHERD). One additional difficulty stems from the fact that the *graphemes*, the orthographic counterparts of phonemes, can consist either of single letters or of letter groups, as the previous examples illustrate.

Psycholinguistic theories of visual word recognition have taken the quasi-systematicity of writing into account in two opposite ways. In one framework, generally known as dual-route theories (e.g. Coltheart, 1978; Coltheart, Curtis, Atkins, & Haller, 1993), it is assumed that dominant mapping regularities are abstracted to derive a tabulation of grapheme-phoneme correspondence rules, which may then be looked up to derive a pronunciation for any letter string. Because the rule table only captures the dominant regularities, it needs to be complemented by lexical knowledge to handle deviations and ambiguities (i.e., CHAOS, SHEPHERD). The opposite view, based on the parallel distributed processing framework, assumes that the whole set of grapho-phonological regularities is captured through differentially weighted associations between letter coding and phoneme coding units of varying sizes (Seidenberg & McClelland, 1989; Plaut, Seidenberg, McClelland & Patterson, 1996).

These opposing theories have nourished an ongoing complex empirical debate for a number of years. This controversy constitutes one instance of a more general issue in cognitive science, which bears upon the proper explanation of rule-like behavior. Is the language user's capacity to exploit print-sound regularities, for instance to generate a plausible pronunciation for a new, unfamiliar string of letters, best explained by knowledge of abstract all-or-none rules, or of the

statistical structure of the language? We believe that, in the field of visual word processing, the lack of precise quantitative descriptions of the mapping system is one factor that has impeded resolution of these issues.

In this paper, we present a descriptive analysis of the grapheme-phoneme mapping system of the French orthography, and we further explore the sensitivity of adult human readers to some characteristics of this mapping. The results indicate that human naming performance is influenced by the frequency of graphemic units in the language and by the predictability of their mapping to phonemes. We argue that these results implicate the availability of graded knowledge of grapheme-phoneme mappings and hence, that they are more consistent with a parallel distributed approach than with the abstract rules hypothesis.

1. Statistical analysis of grapho-phonological correspondences of French

1.1. Method

Tables of grapheme-phoneme associations (henceforth, GPA) were derived from a corpus of 18.510 French one-to-three-syllable words from the BRULEX Database (Content, Mousty, & Radeau, 1990), which contains orthographic and phonological forms as well as word frequency statistics. As noted above, given that graphemes may consist of several letters, the segmentation of letter strings into graphemic units is a non-trivial operation. A semi-automatic procedure similar to the rule-learning algorithm developed by Coltheart *et al.* (1993) was used to parse words into graphemes.

First, grapheme-phoneme associations are tabulated for all trivial cases, that is, words which have exactly the same number of graphemes and phonemes (i.e. PAR, /paʀ/). Then a segmentation algorithm is applied to the remaining unparsed words in successive passes. The aim is to select words for which the addition of a single new GPA would resolve the parsing. After each pass, the new hypothesized associations are manually checked before inclusion in the GPA table.

The segmentation algorithm proceeds as follows. Each unparsed word in the corpus is scanned from left to right, starting with larger letter groups, in order to find a parsing based on tabulated GPAs which satisfies the phonology. If this fails, a new GPA will be hypothesized if there is only one unassigned letter group and one unassigned phoneme and their positions match. For instance, the single-letter grapheme-phoneme associations tabulated at the initial stage would be used to mark the P-/p/ and R-/ʀ/ correspondences in the word POUR (/puʀ/) and isolate OU-/u/ as a new plausible association.

When all words were parsed into graphemes, a

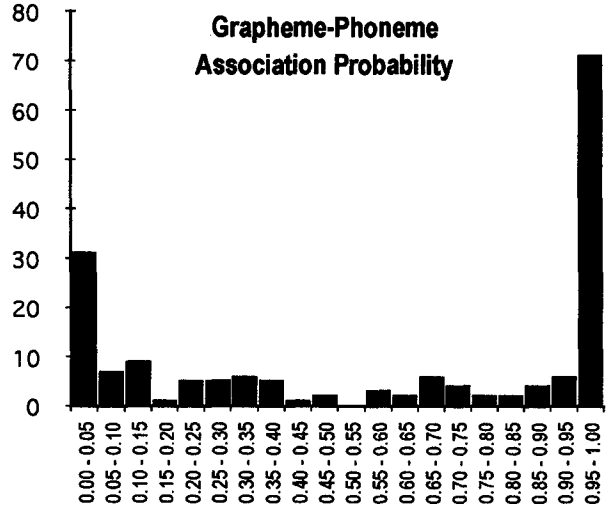

Figure 1. Distribution of Grapheme-Phoneme Association probablity, based on type measures.

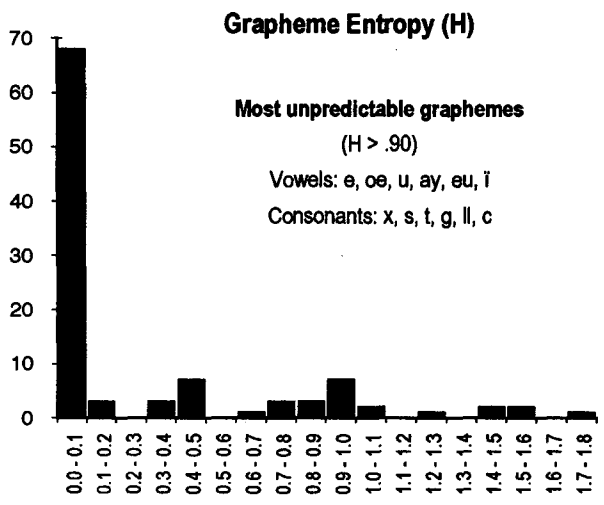

Figure 2. Distribution of Grapheme Entropy (H) values, based on type measures.

	Predictibility of Grapheme-Phoneme Associations in French									
	Number of pronunciations		GPA probability (type)		GPA probability (token)		H (type)		H (token)	
	M	SD	M	SD	M	SD	M	SD	M	SD
All	1.70	(1.26)	.60	(.42)	.60	(.43)	.27	(.45)	.23	(.42)
Vowels	1.66	(1.12)	.60	(.41)	.60	(.44)	.29	(.48)	.21	(.41)
Consonants	1.76	(1.23)	.60	(.42)	.60	(.42)	.25	(.42)	.26	(.44)

Table 1. Number of different pronunciations of a grapheme, grapheme-phoneme association (GPA) probability, and entropy (H) values, by type and by token, for French polysyllabic words.

final pass through the whole corpus computed grapheme-phoneme association frequencies, based both on a type count (the number of words containing a given GPA) and a token count (the number of words weighted by word frequency).

Several statistics were then extracted to provide a quantitative description of the grapheme-phoneme system of French. (1) *Grapheme frequency*, the number of occurrences of the grapheme in the corpus, independently of its phonological value. (2) *Number of alternative pronunciations* for each grapheme. (3) *Grapheme entropy* as measured by *H*, the information statistic proposed by Shannon (1948) and previously used by Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty (1995). This measure is based on the probability distribution of the phoneme set for a given grapheme and reflects the degree of predictability of its pronunciation. *H* is minimal and equals 0 when a grapheme is invariably associated to one phoneme (as for J and /ʒ/). *H* is maximal and equals $log_2\ n$ when there is total uncertainty. In this particular case, *n* would correspond to the total number of phonemes in the language (thus, since there are 46 phonemes, max $H = 5.52$). (4) *Grapheme-phoneme association probability*, which is the GPA frequency divided by the total grapheme frequency. (5) *Association dominance rank*, which is the rank of a given grapheme-phoneme association among the phonemic alternatives for a grapheme, ordered by decreasing probability.

1.2. Results

Despite its well-known complexity and ambiguity in the transcoding from sound to spelling, the French orthography is generally claimed to be very systematic in the reverse conversion of spelling to sound. The latter claim is confirmed by the present analysis. The grapheme-phoneme associations system of French is globally quite predictable. The GPA table includes 103 graphemes and 172 associations, and the mean association probability is relatively high (i.e., 0.60). Furthermore, a look at the distribution of grapheme-phoneme association probabilities (Figure 1) reveals that more than 40% of the associations are completely regular and unambiguous. When multiple pronunciations exist (on average, 1.70 pronunciations for a grapheme), the alternative pronunciations are generally characterized by low GPA probability values (i.e., below 0.15).

The predictability of GPAs is confirmed by a very low mean entropy value. The mean entropy value for all graphemes is 0.27. As a comparison point, if each grapheme in the set was associated with two phonemes with probabilities of 0.95 and 0.05, the mean H value would be 0.29. There is no notable difference between vowel and consonant predictability. Finally, it is worth noting that in general, the descriptive statistics are similar for type and token counts.

2. Empirical study: Grapheme frequency and grapheme entropy

To assess readers' sensitivity to grapheme frequency and grapheme entropy we collected naming latencies for pseudowords contrasted on those two dimensions.

	Grapheme Frequency		Grapheme Entropy	
	Low	High	Low	High
Latencies				
Immediate Naming	609 (75)	585 (66)	596 (72)	644 (93)
Delayed Naming	335 (42)	342 (53)	333 (51)	360 (54)
Delta Scores	274 (94)	243 (84)	263 (94)	284 (105)
Errors				
Immediate Naming	8.1 (7.0)	8.9 (5.8)	9.2 (4.7)	14.2 (7.3)
Delayed Naming	2.7 (3.4)	3.9 (5.7)	2.5 (2.4)	8.0 (6.3)

Table 2. Average reaction times and errors for the grapheme frequency and grapheme entropy (uncertainty) manipulations (standard deviations are indicated into parentheses) in the immediate and delayed naming tasks.

2.1. Method

Participants. Twenty French-speaking students from the Free University of Brussels took part in the experiment for course credits. All had normal or corrected to normal vision.

Materials. Two lists of 64 pseudowords were constructed. The first list contrasted grapheme frequency and the second manipulated grapheme entropy. The grapheme frequency and grapheme entropy estimates for pseudowords were computed by averaging respectively grapheme frequency or grapheme entropy across all graphemes in the letter string. Low and high values items were selected among the lowest 30% and highest 30% values in a database of about 15.000 pseudowords constructed by combining phonotactically legal consonant and vocalic clusters.

The frequency list comprised 32 pairs of items. In each pair, one pseudoword had a high averaged grapheme frequency, and the other had a low averaged grapheme frequency, with entropy kept constant. Similarly, the entropy list included 32 pairs of pseudowords with contrasting average values of entropy and close values of average grapheme frequency.

In addition, stimuli in a matched pair were controlled for a number of orthographic properties known to influence naming latency (number of letters and phonemes; lexical neighborhood size; number of body friends; positional and non positional bigram frequency; grapheme segmentation probability; grapheme complexity).

Procedure. Participants were tested individually in a computerized situation (PC and MEL experimentation software). They were successively tested in a immediate naming and a delayed naming task with the same stimuli. In the immediate naming condition, participants were instructed to read aloud pseudowords as quickly and as accurately as possible, and we recorded response times and errors. In the delayed naming task, the same stimuli were presented in a different random order, but participants were required to delay their overt response until a response signal appeared on screen. The delay varied randomly from trial to trial between 1200 and 1500 msec. Since participants are instructed to fully prepare their response for overt pronunciation during the delay period, the delayed naming procedure is meant to provide an estimate of potential artefactual differences between stimulus sets due to articulatory factors and to differential sensitivity of the microphone to various onset phonemes.

Pseudowords were presented in a random order, different for each participant, with a pause after blocks of 32 stimuli. They were displayed in lower case, in white on a black background. In the immediate naming task, each trial began with a fixation sign (*) presented at the center of the screen for 300 msec. It was followed by a black screen for 200 msec and then a pseudoword which stayed on the screen until the vocal response triggered the microphone or for a maximum delay of 2000 msec. An interstimulus screen was finally presented for 1000 msec. In the delayed naming task, the fixation point and the black screen were

followed by a pseudoword presented for 1500 msec, followed by a random delay between 1300 and 1500 msec. After this variable delay, a go signal (####) was displayed in the center of the screen till a vocal response triggered the microphone or for a maximum duration of 2000 msec. Pronunciation errors, hesitations and triggering of the microphone by extraneous noises were noted by hand by the experimenter during the experiment.

2.2. Results

Data associated with inappropriate triggering of the microphone were discarded from the error analyses. In addition, for the response time analyses, pronunciation errors, hesitations, and anticipations in the delayed naming task were eliminated. Latencies outside an interval of two standard deviations above and below the mean by subject and condition were replaced by the corresponding mean. Average reaction times and error rates were then computed by subjects and by items in both the immediate naming and the delayed naming task. By-subjects and by-items (F_1 and F_2, respectively) analyses of variance were performed with grapheme frequency and grapheme entropy as within-subject factors.

Grapheme frequency. For naming latencies, pseudowords of low grapheme frequency were read 24 msec more slowly than pseudowords of high grapheme frequency. This difference was highly significant both by subjects and by items; $F_1(1, 19) = 24.4, p < .001$, $F_2(1, 31) = 7.5, p < .001$. On delayed naming times, the same comparison gave a nonsignificant difference of -7 msec. For pronunciation errors, there was no significant difference in the immediate naming task. In the delayed naming task, pseudowords of low mean grapheme frequency caused 1.2% more errors than high ones. This difference was marginally significant by items, but not significant by subjects; $F_2(1, 31) = 3.1, p < .1$.

Grapheme entropy. In the immediate naming task, high-entropy pseudowords were read 48 msec slower than low-entropy pseudowords; $F_1(1, 19) = 45.4, p < .001$, $F_2(1, 31) = 16.2, p < .001$. In the delayed naming task, the same comparison showed a significant difference of 27 msec; $F_1(1, 19) = 22.9\ p < .001$, $F_2(1, 31) = 12.5, p < .005$. Because of this articulatory effect, delta scores were computed by subtracting delayed naming times from immediate naming times. A significant difference of 21 msec was found on delta scores; $F_1(1, 19) = 5.7, p < .05$, $F_2(1, 31) = 4.7, p < .05$.

The pattern of results was similar for errors. In the immediate naming task, high-entropy pseudowords caused 5% more errors than low-entropy pseudowords. This effect was significant by subjects but not by items; $F_1(1, 19) = 7.4, p < .05$, $F_2(1, 31) = 2.1, p > .1$. The effect was of 6.5% in the delayed naming task and was significant by subjects and items; $F_1(1, 19) = 17.2, p < .001$, $F_2(1, 31) = 8.3, p < .01$.

2.3. Discussion

A clear effect of the grapheme frequency and the grapheme entropy manipulations were obtained on immediate naming latencies. In both manipulations, the stimuli in the contrasted lists were selected pairwise to be as equivalent as possible in terms of potentially important variables.

A difference between high and low-entropy pseudowords was also observed in the delayed naming condition. The latter effect is probably due to phonetic characteristics of the initial consonants in the stimuli. Some evidence confirming this interpretation is adduced from a further control experiment in which participants were required to repeat the same stimuli presented auditorily, after a variable response delay. The 27 msec difference in the visual delayed naming condition was tightly reproduced with auditory stimuli, indicating that the effect in the delayed naming condition is unrelated to print-to-sound conversion processes. Despite this unexpected bias, however, when the influence of phonetic factors was eliminated by computing the difference between immediate and delayed naming, a significant effect of 21 msec remained, demonstrating that entropy affects grapheme-phoneme conversion.

These findings are incompatible with current implementations of the dual-route theory (Coltheart *et al.*, 1993). The "central dogma" of this theory is that the performance of human subjects on pseudowords is accounted for by an analytic process based on grapheme-phoneme conversion rules. Both findings are at odds with the additional core assumptions that (1) *only*

dominant mappings are retained as conversion rules; (2) there is no place for ambiguity or predictability in the conversion.

In a recent paper, Rastle and Coltheart (1999) note that "One refinement of dual-route modeling that goes beyond DRC in its current form is the idea that different GPC rules might have different strengths, with the strength of the correspondence being a function of, for example, the proportion of words in which the correspondence occurs. Although simple to implement, we have not explored the notion of rule strength in the DRC model because we are not aware of any work which demonstrates that any kind of rule-strength variable has effects on naming latencies when other variables known to affect such latencies such as neighborhood size (e.g., Andrews, 1992) and string length (e.g., Weekes, 1997) are controlled."

We believe that the present results provide the evidence that was called for and should incite dual-route modelers to abandon the idea of all-or-none rules which was a central theoretical assumption of these models compared to connectionist ones. As the DRC model is largely based on the interactive activation principles, the most natural way to account for graded effects of grapheme frequency and pronunciation predictability would be to introduce grapheme and phoneme units in the nonlexical system. Variations in the activation resting level of grapheme detectors as a function of frequency of occurrence and differences in the strength of the connections between graphemes and phonemes as a function of association probability would then explain grapheme frequency and grapheme entropy effects. However an implementation of rule-strength in the conversion system of the kind suggested considerably modifies its processing mechanism, notably by replacing the serial table look-up selection of graphemes by a parallel activation process. Such a change is highly likely to induce non-trivial consequences on predicted performance.

Furthermore, and contrary to the suggestion that the introduction of rule-strength would amount to a mere implementational adaptation of no theoretical importance, we consider that it would impose a substantial restatement of the theory, because it violates the core assumption of the approach, namely, that language users induce all-or-none rules from the language to which they are exposed. Hence, the cost of such a (potential) improvement in descriptive adequacy is the loss of explanatory value from a psycholinguistic perspective. As Seidenberg stated, "[we are] not claiming that data of the sort presented [here] cannot in principle be accommodated within a dual route type of model. In the absence of any constraints on the introduction of new pathways or recognition processes, models in the dual route framework can always be adapted to fit the empirical data. Although specific proposals might be refuted on the basis of empirical data, the general approach cannot." (Seidenberg, 1985, p. 244).

The difficulty to account for the present findings within the dual-route approach contrasts with the straigthforward explanation they receive in the PDP framework. As has often been emphasized, rule-strength effects emerge as a natural consequence of learning and processing mechanisms in parallel distributed systems (see Van Orden, Pennington, & Stone, 1990; Plaut *et al.*, 1996). In this framework, the rule-governed behavior is explained by the gradual encoding of the statistical structure that governs the mapping between orthography and phonology.

Conclusions

In this paper, we presented a semi-automatic procedure to segment words into graphemes and tabulate grapheme-phoneme mappings characteristics for the French writing system. In current work, the same method has been applied on French and English materials, allowing to provide more detailed descriptions of the similarities and differences between the two languages. Most previous work in French (e.g. Véronis, 1986) and English (Venezky, 1970) has focused mainly on the extraction of a rule set. One important feature of our endeavor is the extraction of several quantitative graded measures of grapheme-phoneme mappings (see also Berndt, Reggia, & Mitchum, 1987, for similar work in American English).

In the empirical investigation, we have shown how the descriptive data could be used to probe human readers' written word processing. The results demonstrate that the descriptive statistics

capture some important features of the processing system and thus provide an empirical validation of the approach. Most interestingly, the sensitivity of human processing to the *degree* of regularity and frequency of grapheme-phoneme associations provides a new argument in favor of models in which knowledge of print-to-sound mapping is based on a large set of graded associations rather than on correspondence rules.

Acknowledgements

This research was supported by a research grant from the Direction Générale de la Recherche Scientifique — Communauté française de Belgique (ARC 96/01-203). Marielle Lange is a research assistant at the Belgian National Fund for Scientific Research (FNRS).

References

Andrews, S. (1992). *Frequency and neighborhood effects on lexical access: Lexical similarity or orthographic redundancy?* Journal of Experimental Psychology: Learning, Memory, and Cognition, 18, 234-254.

Berndt, R. S., Reggia, J. A., & Mitchum, C. C. (1987). *Empirically derived probabilities for grapheme-to-phoneme correspondences in English.* Behavior Research Methods, Instruments, & Computers, 19, 1-9.

Chater, N., & Christiansen, M. H. (1998). *Connectionism and Natural Language Processing.* In S. Garrod & M. Pickering. (Eds.), Language Processing. London, UK: University College London Press.

Coltheart, M. (1978). *Lexical access in simple reading tasks.* In G. Underwood (Ed.), Strategies of information processing (pp. 151-216). London: Academic Press.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). *Models of reading aloud: Dual-route and parallel-distributed-processing approaches.* Psychological Review, 100, 589-608.

Content, A., Mousty, P., & Radeau, M. (1990). *Brulex. Une base de données lexicales informatisée pour le français écrit et parlé* [Brulex, A lexical database for written and spoken French]. L'Année Psychologique, 90, 551-566.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). *Understanding normal and impaired word reading: Computational principles in quasi-regular domains.* Psychological Review, 103, 56-115.

Rastle, K., & Coltheart, M. (1999). *Serial and strategic effects in reading aloud.* Journal of Experimental Psychology: Human Perception and Performance, (April, 1999, in press).

Seidenberg, M. S. (1985). *The time course of information activation and utilization in visual word recognition.* In D. Besner, T. G. Waller, & E. M. MacKinnon (Eds.), Reading Research: Advances in theory and practice (Vol. 5, pp. 199-252). New York: Academic Press.

Seidenberg, M. S., & McClelland, J. L. (1989). *A distributed, developmental model of word recognition and naming.* Psychological Review, 96, 523-568.

Shannon, C. E. (1948). *A mathematical theory of communication.* Bell System Technical Journal, 27, 379-423, 623-656.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E. D. (1995). *The special role for rimes in the description, use, and acquisition of English Orthography.* Journal of Experimental Psychology: General, 124, 107-136.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). *Word identification in reading and the promise of subsymbolic psycholinguistics.* Psychological Review, 97, 488-522.

Venezky, R. L. (1970). *The structure of English orthography.* The Hage, The Netherlands: Mouton.

Véronis, J. (1986). *Etude quantitative sur le système graphique et phonologique du français.* Cahiers de Psychologie Cognitive, 6, 501-531.

Weekes, B. (1997). *Differential effects of letter number on word and nonword naming latency.* Quarterly Journal of Experimental Psychology, 50A, 439-456.