Pronouncing Text by Analogy

Robert I. Damper and John F.G. Eastmond
Image, Speech and Intelligent Systems (ISIS) Research Group,
Department of Electronics and Computer Science,
University of Southampton,
Southampton SO17 1BJ,
UK.
{rid|je}@ecs.soton.ac.uk

Abstract

Pronunciation-by-analogy (PbA) is an emerging technique for text-phoneme conversion based on a psychological model of reading aloud. This paper explores the impact of certain basic implementational choices on the performance of various PbA models. These have been tested on their ability to pronounce sets of short pseudowords previously used in similar studies, as well as lexical words temporarily removed from the dictionary. Best results of 85.7% and 67.9% words correct are obtained for the pseudowords and lexical words respectively, casting doubt on certain previous-reported performance figures in the literature.

1 Introduction

Pronunciation-by-analogy (PbA) is an influential psychological model of the process of reading aloud. In PbA, most words are pronounced by retrieving their phonemic form from the readers's lexicon, or dictionary. The pronunciation for a 'novel' word not in the lexicon, however, is derived not by the application of abstract letter-to-sound rules but is 'assembled' from the (known) pronunciations of words that it resembles. PbA has obvious application to text-to-speech conversion by machine.

Although PbA programs have been presented in the literature, they are they are few in number. Dedina and Nusbaum (1991) describe PRONOUNCE: a rather simple system for English. Sullivan and Damper (1990; 1992; 1993) describe a considerably more complex and developed system, but which apparently yields a much poorer performance.

As a psychological theory, PbA is under-specified: offering little meaningful guidance on the implementation choices which confront the programmer. Indeed, Sullivan and Damper (1993) show that such choices can have a profound impact on performance. In this paper, we seek to understand how Dedina and Nusbaum's largely unjustified implementational choices affected their results and, thereby, to resolve the conflict between their performance claims and Sullivan and Damper's.

2 Psychological Background

In the standard *dual-route* model of reading aloud (Coltheart, 1978), there is a lexical route for the pronunciation of known words and a parallel route utilising abstract letter-to-sound rules for the pronunciation of unknown ('novel') words. Arguments for dual-route theory cite the ability to pronounce pseudowords (non-words conforming to the spelling patterns of English), latency difference effects between regular and exception words, and apparent double dissociation between the two routes in dyslexia (see Humphreys and Evett, 1985). However, all these observations can arguably be explained by a single route. One pervasive idea is that pseudowords are pronounced by analogy with lexical words that they resemble (Baron, 1977; Brooks, 1977; Glushko, 1979; 1981; Brown and Besner, 1987). Glushko, for instance, showed that "exception pseudowords" like *tave* take longer to read than "regular pseudowords" such as *taze*. Here, *taze* is considered as a "regular pseudoword" since all its orthographic 'neighbours' (*raze, gaze, maze* etc.) have the regular vowel pronunciation /eI/. By contrast, *tave* is considered to be an "exception pseudoword" since it has the exception word (*have*, /hav/) as an orthographic neighbour. Thus, according to Glushko (1979), the "assignment of phonology to non-words is open to lexical influence". This is at variance with the notion of two independent routes to pronunciation. Instead:

> "it appears that words and pseudowords are pronounced using similar kinds of orthographic and phonological knowledge: the pronunciation of words that share orthographic features with them, and specific spelling-to-sound rules for multiletter spelling patterns."

There are two forms of PbA: *explicit* analogy (Baron, 1977) is a conscious strategy of recalling a similar word and modifying its pronunciation, whereas in *implicit* analogy (Brooks, 1977) a pronunciation is derived from generalised phonographic knowledge about existing words. The latter has obvious commonalities with most single-route, connectionist models (e.g. Sejnowski and Rosenberg, 1987) in which the generalised knowledge is learned (e.g. by back-propagation) as a set of weights, and the network has no holistic notion of the concept 'word'.

Until the recent advent of computational PbA models, analogy 'theory' could only be considered seriously underspecified. Clearly, its operation must depend critically on some measure of similarity, and "without a metric for similarity and without a specification of how similar is similar enough, the concept of analogy by similarity offers little insight" (Glushko, 1981, p. 72). Further, as detailed by Brown and Besner (1987), the operation of lexical analogy must be constrained by factors such as:

- the size of the segment shared between novel and lexical word;
- its position in the two strings;
- its frequency of occurrence in the language;
- and the frequency of occurrence of the words containing it;

none of which had then received serious consideration. Accordingly, they write: "Extant analogy models are not capable of predicting the outcome of assembly operations for all possible strings."

In particular, the 'theory' gives no principled way of deciding the orthographic neighbours of a novel word which are deemed to influence its pronunciation whereas a computational model must (specifically or otherwise) do so.

3 Existing PbA Programs

3.1 Dedina and Nusbaum's System

The overall structure of PRONOUNCE is as shown in Fig. 1. The lexical database consists of "approximately 20,000 words based on *Webster's Pocket Dictionary*" in which text and phonemes have been automatically aligned. Dedina and Nusbaum acknowledge the crude nature of their alignment procedure, saying it "was carried out by a simple Lisp program that only uses knowledge about which phonemes are consonants and which are vowels."

An input string is matched in turn against all orthographic entries in the lexicon. The process starts with the input string and the current dictionary entry left-aligned. Information about matching letter substrings

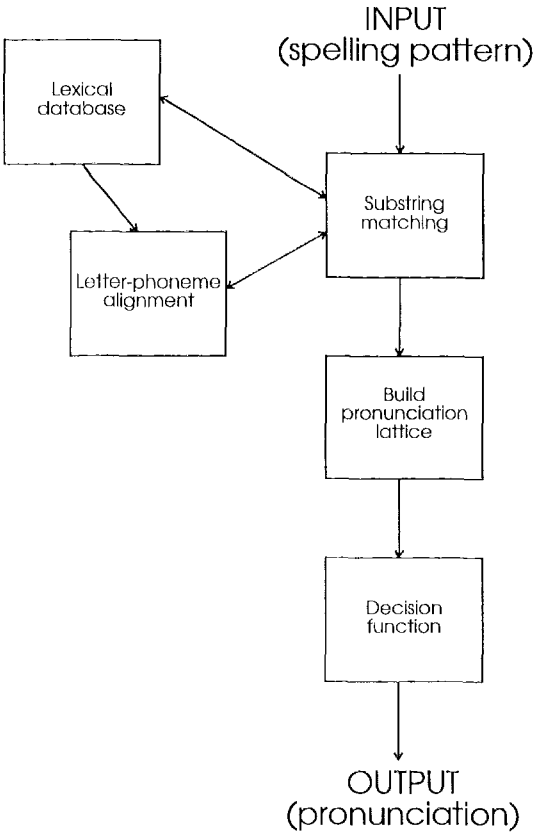

Figure 1: Dedina and Nusbaum's PRONOUNCE.

– and their corresponding phoneme substrings in the dictionary entry under consideration – is entered into a *pronunciation lattice* as detailed below. The shorter of the two strings is then shifted right by one letter and the process repeated. This continues until the two are right-aligned, i.e. the number of right shifts is equal to the difference in length between the two strings. The process is repeated for all words in the dictionary.

A node of the lattice represents a matched letter, L_i, at some position, i, in the input, as illustrated in Fig. 2. The node is labelled with its position index i and with the phoneme which corresponds to L_i in the matched substring, P_{im} say, for the mth matched substring. An arc is placed from node i to node j if there is a matched substring starting with L_i and ending with L_j. The arc is labelled with the phonemes intermediate between P_{im} and P_{jm} in the phoneme part of the matched substring. Note that the empty string labels arcs corresponding to bigrams: the two symbols of the bigram label the nodes at either end. Additionally, arcs are labelled with a "frequency" count which is incremented by one each time that substring (with that pronunciation) is matched during the pass through the lexicon. Finally, there is a *Start* node at position 0 and an *End* node at position one greater than the length of the input string.

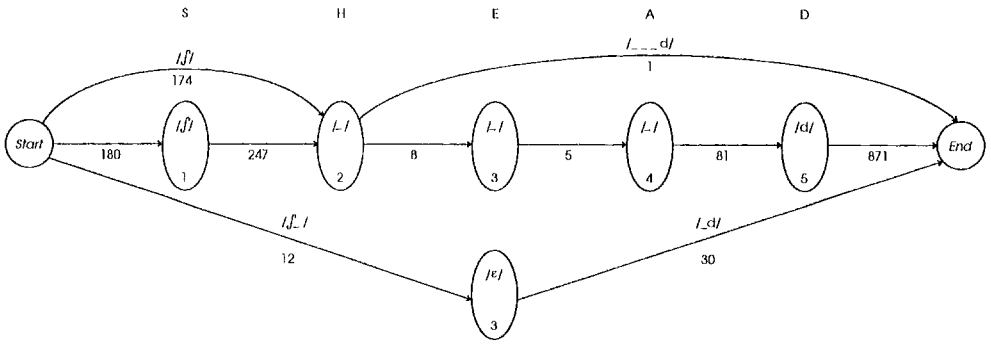

Figure 2: Partial pronunciation lattice for the pseudoword *shead*.

A possible pronunciation for the input corresponds to a complete path through its lattice from *Start* to *End*, with the output string assembled by concatenating in order the phoneme labels on the nodes/arcs. The set of candidate pronunciations is then passed to the decision function. Two (prioritised) heuristics are used to rank the pronunciations, and the top-ranking candidate selected as the output. The first is based on path length. If one candidate corresponds to a unique shortest path (in terms of number of arcs) through the lattice, this is selected as the output. Otherwise, candidates that tie are ranked on the sum of their arc "frequencies".

Dedina and Nusbaum tested PRONOUNCE on 70 of Glushko's (1979) pseudowords, which "were four or five characters long and were derived from monosyllabic words by changing one letter". Seven subjects with phonetics training were asked to read these and give a transcription for the first pronunciation which came to mind. A 'correct' pronunciation for a given pseudoword was considered to be one produced by any of the subjects. A word error rate of 9% is reported.

3.2 Sullivan and Damper's System

Sullivan and Damper employ a more principled alignment procedure based on the Lawrence and Kaye (1986) algorithm. By pre-computing mappings and their statistics, they implemented a considerably more 'implicit' form of PbA: there is no explicit matching of the input string with lexical entries. Their pronunciation lattice differs, with nodes representing *junctures* between symbols and arcs representing letter-phoneme mappings. They also examine different ways of numerically ranking candidates, taking into account probabilities estimates for the letter-phoneme mappings used in the assembled pronunciation.

Given the improved alignment and candidate-ranking methods, better performance than Dedina and Nusbaum might be expected. On the contrary, Sullivan and Damper's best result on the full set of 131 pseudowords from Glushko (1979) (plus another 5 words -- see section 5.1) is only 70.6% (1993, p. 449). This is an error rate of almost 30%, as compared to Dedina and Nusbaum's 9% on the smaller test set of size 70. Differences in test-set size and between British and American English, the transcription standards of the phoneticians, and the lexicons employed seem insufficient to explain this.

4 Re-Implementing PRONOUNCE

Our purpose was to re-implement PRONOUNCE, assess its performance, and study the impact of various implementational choices on this performance. However, the described alignment algorithm is problematic (see pp. 71–73 of Sullivan, 1992) and needs to be replaced. Rather than re-implement a flawed algorithm, we have used manually-aligned data. Since manual alignment generally produces a better result than automatic alignment, we ought to produce an *even* lower error rate than Dedina and Nusbaum's claimed 9%.

The performance on lexical words (temporarily removed from the lexicon) has not previously been assessed but seems worthwhile. Arguably, 'real' words form a much more sensible test set for a PbA system than pseudowords, not least because they are multisyllabic. Temporary removal from the lexicon means that the pronunciation must be assembled by the analogy process rather that merely retrieved in its entirety. Hence, we believe it is sensible and important to test any PbA system in this way.

4.1 Lexical Databases

To examine any impact that the specific lexical database might have on performance, we have used two in this work: the 20,009 words of *Webster's Pocket Dictionary* and the 16,280 words of the *Teacher's Word Book* (TWB) (Thorndike and Lorge, 1944). In both cases, letters and phonemes have previously been hand-aligned for the purposes of training back-propagation networks. The Webster's database is that used by Sejnowski and Rosenberg (1987) to train and test NETtalk. The TWB database is that used by McCulloch, Bedworth and Bridle (1987) for NETspeak.

The phoneme inventory is of size 52 in both cases, including the null phoneme but excluding stress symbols. We leave the very important problem of stress assignment for later study.

4.2 Re-Implementation Details

The re-implementation was programmed in C on a Hewlett-Packard 712/80 workstation running HP-UX. A 'direct' version scores candidates using Dedina and Nusbaum's method with its two prioritised heuristics: we call this model D&N. Two other methods for scoring have also been implemented. In one, we replace the second (maximum sum) heuristic with the maximum product of the arc frequencies: we call this model PROD. (It still selects primarily on the basis of shortest path length.) We have also implemented a version which uses a single heuristic. This takes the product along each possible path from *Start* to *End* of the *mapping probabilities* for that arc. These are computed using Method 1 (*a priori* version) of Sullivan and Damper (1993, pp. 446–447). For all paths corresponding to the *same* pronunciation, these values are summed to give an overall score for that pronunciation. We call this the MP model. The final product score is *not* a proper probability for the assembled pronunciation, since the scores do not sum to one over all the candidates.

The 'best' pronunciation is found by depth-first search of the lattice, implemented as a preorder tree traversal. For the D&N and PROD models, paths were pruned when their length exceeded the shortest found so far for that input, leading to a useful reduction in run times. A similarly motivated pruning was carried out for the MP model. If any product fell below a threshold during traversal, its corresponding path was discarded. The threshold used was ϵ times the maximum product score found so far, with ϵ set by at 10^{-3}. While this may have led to the pruning of a path contributing to the 'best' pronunciation, its contribution would be very small. Again, this gave a very significant improvement in run times for the testing of lexical words (section 5.2 below) but was unnecessary for the testing of pseudowords.

5 Results

5.1 Pseudowords

Pronunciations have been obtained for:

- the 70 pseudowords from Glushko (1979) used by Dedina and Nusbaum to test PRONOUNCE. The 'correct' pronunciation for these strings is taken to be that given by Dedina and Nusbaum (1991, pp. 61–62). We refer to this test set as D&N 70.
- the full set of 131 pseudowords from Glushko plus two others (*goot*, *pome*) plus two lexical words (*cat* and *play*) plus the pseudohomophone *kwik*, as used by Sullivan (1992). The 'correct' pronunciations are those read aloud by Sullivan's 20 non-phonetician subjects, and transcribed by him as British Received Pronunciation. We refer to this test set as Sull 136. Our expectation is that the error rate will be relatively high for this test set, partly because of its larger size but more importantly because the subjects' dialect of English is British RP rather that general American, i.e. there is a very significant inconsistency with the lexical databases.

The output has been scored on words correct and also on symbol score (i.e. phonemes correct) using the Levenshtein (1966) string-edit distance as shown in Table 1.

Our best comparison with Dedina and Nusbaum (D&N 70 test set, D&N model, Webster's database) gives a figure of 77.1% words correct. This is enormously poorer than their approximately 91% words correct – yet the implementation, reference pronunciations and test set are (as far as we can tell) identical. The only relevant difference is that the Webster's database is automatically-aligned in their work and hand-aligned in ours. The clear expectation, given the crude nature of their alignment, is that they should have experienced a *higher* error rate, not a dramatically lower one. Overall, this result accords far more closely with Sullivan and Damper (1993) whose best word score for automatic alignment (and using smaller databases but a larger test set) was just over 70%.

The re-implementation made 16 errors under the above conditions. Dedina and Nusbaum's claim of 9% words correct amounts to just 6 errors, 3 of which are the same as ours. The commonest problem is vowel substitution. It is possible to discount a very few errors as essentially trivial, reducing the error rate marginally to some 20%. We conclude, therefore, that Dedina and Nusbaum's reported error rate of 9% is unattainable.

In our opinion, a major deficiency of the simple shortest-path length heuristic is that the output can become unreasonably sensitive to rare or unique pronunciations. For instance, *mone* receives the strange pronunciation /məni/ by analogy with *anemone*. Also, the pseudoword *shead* receives the bizarre, vowelless pronunciation /ʃ_ _ _d/ (where '_' denotes the null phoneme) when using the D&N model and the TWB database. As illustrated in Fig. 2 earlier, this turns out to be a result of matching the unique but long mapping *head* → /_ _ _d/ as in *forehead* → /fɔr_ _ _ _d/ (arc frequency 1) in conjunction with the very common mapping *sh* → /ʃ_/ as in *she* and *shed* (arc frequency 174) which swamps the overall score of 175. The same bizarre pronunciation does not occur with the PROD model. In this case, the path through the

Table 1: Results for PbA of pseudowords with both dictionaries. See text for further specification.

Test set	Implementation	Webster's (%)		TWB (%)	
		words	phonemes	words	phonemes
D&N 70	D&N	77.1	94.3	70.0	92.6
	PROD	82.9	95.9	78.6	94.9
	MP	85.7	96.6	80.0	95.3
Sull 136	D&N	75.0	93.6	72.1	93.1
	PROD	80.1	95.0	76.5	94.5
	MP	83.8	95.9	81.6	95.7

(/ɛ/, 3) node has a product score of $12 \times 30 = 360$ for the pronunciation /ʃɛd/ which considerably exceeds the score of 174 for /ʃd/.

Replacing the arc-sum heuristic of the D&N model by arc-product as in the PROD model leads to a considerable increase in performance, e.g. from 77.1% words correct to 82.9% for the D&N 70 test set with Webster's database. In turn, the MP model performs better than PROD in all cases.

For the Sull 136 test set, our expectation of poorer performance (because of the larger test set and inconsistency between of dialect between the target pronunciations and the lexical databases) is borne out for Webster's dictionary. For TWB, however, the performance difference between test sets is less consistent.

5.2 Lexical Words

The primary ability of a text-to-speech system must be to produce correct pronunciations for lexical words (rather than pseudowords) which just happen to be absent from the system's dictionary. Accordingly, we have tested the PbA implementations by removing each word in turn from its relevant database, and obtaining a pronunciation by analogy with the remainder. In these tests, the transcription standard employed by the compilers of the dictionary becomes its own reference and problems of transcription inconsistencies between input strings and lexical entries are avoided.

Results for the testing of lexical words are shown in Table 2. Again there are consistent performance differences with the 'standard' D&N model worst and the mapping probability (MP) model best. All models perform better with the TWB database than with Webster's, probably simply because of its smaller size.

For some lexical words, no pronunciation at all was produced because there was no complete path from *Start* to *End* in the lattice. This occurred for 92 of the TWB words and 117 of the Webster's words irrespective of the scoring model. This is a serious shortcoming: a PbA system should always produce a best-attempt pronunciation, even if it cannot produce the correct one. Sometimes, this failure is a consequence of the form of pronunciation lattice in which nodes are used to represent the 'end-points' of mappings. One of the inputs for which no pronunciation was found is *anecdote*, whose (partial) lattice is shown in Fig. 3. There is in fact no arc in the complete lattice between nodes (/k/, 4) and (/d/, 5) because there is no $cd \rightarrow$ /kd/ mapping anywhere in either dictionary. Nor is there an *ecd* or *cdo* trigram – with or without the right end-point phonemes – which could possibly bridge the gap. This problem is entirely avoided with the Sullivan and Damper style of lattice, because the shortest-length arc corresponds to a single-symbol mapping rather than to a bigram (which may be unique). Thus, there will always be a 'default' single-symbol mapping corresponding to the commonest pronunciation of the letter. This is not to say that Sullivan and Damper's system will necessarily produce the correct output here: it almost certainly will not because of the rarity of the $c \rightarrow$ /k/ mapping in the $\_d$ context.

Another input which fails to produce a pronunciation is *aardvark*. The problem here is not that there is no *aa* bigram in the dictionary (which is found in words such as *bazaar*), but that it only appears towards the end of other words. Dedina and Nusbaum's strategy of performing substring matching only over a restricted range (the number of matching comparisons is equal to the difference in length between the input string and lexical entry) is at the root of this problem.

6 Conclusions and Discussion

We find that Dedina and Nusbaum's reported error rate of 9% cannot be reproduced: our figure is about two or three times that. Because of the shortcomings which emerge in this work, we believe the problem lies with PRONOUNCE rather than our re-implementation. Overall, our results are in much closer agreement with Sullivan and Damper's word error rates of almost 30% on a similar test set.

This work suggests several useful ways in which the performance of PbA systems might be improved. Our best results are obtained with a scoring method based on *a priori* mapping probabilities. According to Sul-

Table 2: Results for PbA of dictionary words.

Implementation	Webster's (%)		TWB (%)	
	words	phonemes	words	phonemes
D&N	57.8	90.4	65.6	93.1
PROD	58.5	90.7	66.1	93.2
MP	60.7	91.2	67.9	93.5

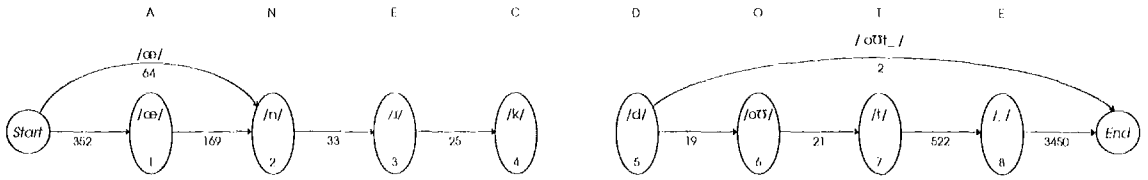

Figure 3: Simplified pronunciation lattice for the lexical word *anecdote* which fails to produce any pronunciation.

livan and Damper (1993), *a posteriori* mapping probabilities may do even better. Also, the type of pronunciation lattice used by Sullivan and Damper, in which nodes correspond to the junctures between symbols, is likely to be superior. The impact of different alignment strategies should repay study. Finally, we intend to assess the impact of incorporating information about word frequency in the analogy process.

Acknowledgement

This work was funded by the UK Economic and Social Research Council via research grant R000235487: "Speech Synthesis by Analogy".

References

Baron, J. (1977). Mechanisms for pronouncing printed words: use and acquisition. In *Basic Processes in Reading: Perception and Comprehension* (D. LaBerge and S. Samuels, eds.), pp. 175–216. Lawrence Erlbaum, Hillsdale, NJ.

Brooks, L. (1977). Non-analytic correspondences and pattern in word pronunciation. In *Attention and Performance VII* (J. Renquin, ed.), pp. 163–177. Lawrence Erlbaum, Hillsdale, NJ.

Brown, P. and Besner, D. (1987). The assembly of phonology in oral reading: a new model. In *Attention and Performance XII: the Psychology of Reading* (M. Coltheart, ed.), pp. 471–489. Lawrence Erlbaum, London.

Coltheart, M. (1978). Lexical access in simple reading tasks. In *Strategies of Information Processing* (G. Underwood, ed.), pp. 151–216. Academic, London.

Dedina, M.J. and Nusbaum, H.C. (1991). PRONOUNCE: a program for pronunciation by analogy. *Computer Speech and Language*, **5**, 55–64.

Glushko, R.J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, **5**, 674–691.

Glushko, R.J. (1981). Principles for pronouncing print: the psychology of phonography. In *Interactive Processes in Reading* (A.M. Lesgold and C.A. Perfetti, eds.), pp. 61–84. Lawrence Erlbaum, Hillsdale, NJ.

Humphreys, G.W. and Evett, L.J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading. *Behavioral and Brain Sciences*, **8**, 689–740.

Lawrence, S.G.C. and Kaye, G. (1986). Alignment of phonemes with their corresponding orthography. *Computer Speech and Language*, **1**, 153–165.

Levenshtein, V.I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Cybernetics and Control Theory*, **10**, 707–710.

McCulloch, N., Bedworth, M. and Bridle, J.S. (1987). NETspeak – a re-implementation of NETtalk. *Computer Speech and Language*, **2**, 289–301.

Sejnowski, T.J. and Rosenberg, C.R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, **1**, 145–152.

Sullivan, K.P.H. (1992). *Synthesis-by-Analogy: a Psychologically Motivated Approach to Text-to-Speech Conversion*, PhD Thesis, Department of Electronics and Computer Science, University of Southampton, UK.

Sullivan, K.P.H. and Damper, R.I. (1990). A psychologically governed approach to novel-word pronunciation within a text-to-speech system. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '90), Vol. 1*, Albuquerque, NM, pp. 341–344.

Sullivan, K.P.H. and Damper, R.I. (1992). Novel-word pronunciation within a text-to-speech system. In *Talking Machines: Theories, Models and Applications* (G. Bailly and C. Benoît, eds.), pp. 183–195. Elsevier (North-Holland), Amsterdam.

Sullivan, K.P.H. and Damper, R.I. (1993). Novel-word pronunciation: a cross-language study. *Speech Communication*, **13**, 441–452.

Thorndike, E.L. and Lorge, I. (1944). *The Teachers' Word Book of 30,000 Words*. Teachers' College, Columbia University, NY.