In P. Broeder & J. Murre (eds.), Models of Language
Acquisition. Oxford University Press, 55–75, 2000

A Distributed, Yet Symbolic Model for Text-to-Speech Processing

Antal van den Bosch
Dept. of Computer Science
Universiteit Maastricht
antal@cs.unimaas.nl

Walter Daelemans
Computational Linguistics
Tilburg University
Walter.Daelemans@kub.nl

Abstract

In this paper, a data-oriented model of text-to-speech processing is described. On the basis of a large text-to-speech corpus, the model automatically gathers a distributed, yet symbolic representation of subword-phoneme association knowledge, representing this knowledge in the form of paths in a decision tree. Paths represent context-sensitive rewrite rules which unambiguously map strings of letters onto single phonemes. The more ambiguous the mapping is, the larger the stored context. The knowledge needed for converting a spelling word to its phonemic transcription is thus represented in a distributed fashion: many different paths contribute to the phonemisation of a word, and a single path may contribute to phonemisations of many words. Some intrinsic properties of the data-oriented model are shown to have relations with psycholinguistic concepts such as a language's orthographic depth, and word pronunciation consistency.

1 Modelling reading aloud

Within psycholinguistics, various models have been proposed in which reading aloud in skilled readers is modelled. These models vary considerably in their restrictions imposed on the representation and processing of knowledge. All of these models face the problem how to explain that experienced human readers are able to pronounce words with regular or irregular pronunciations, and also to pronounce words they have not encountered before (e.g. nonwords). The ability to pronounce nonwords and regular words can be explained in terms of a model which has knowledge of typical spelling-to-phonology rules; the pronunciation of words with irregular pronunciations, however, implies that the model needs some word-specific, lexical knowledge. Furthermore, three general desiderata for a psychological model of skilled reading aloud are: (i) the model is able to account for human performance on the level of word naming reaction times and pronunciation errors, (ii) the model is able to build up its body of knowledge by learning from examples, and

(iii) the model's performance 'gracefully degrades' when it is damaged, or presented with noise.

Before introducing our model, we present three existing examples of models that impose different restrictions on knowledge representation and processing: (i) a *dual-route* model (Coltheart, 1978), in which lexical knowledge and spelling-to-phonology rules are strictly separated, (ii) a *single-route* model (Plaut *et al.*, 1996), in which knowledge is represented and processed in a single system, allowing for any spelling-phonology correspondence ranging from rules to lexical knowledge, and (iii) a *multiple-level processing* model (Norris, 1993), which allows for predefined levels of correspondences between spelling and phonology to interact in one system.

In Coltheart (1978; for an implementation, see also Coltheart *et al.*, 1993), a model is presented which presupposes that there are two separate processing routes available to skilled human readers to convert text to speech: (i) by rule, i.e., converting spelling strings to phonemic strings by applying rules that express grapheme-phoneme correspondences, and (ii) by direct access to a phonological lexicon via the visual recognition of the whole word. The model thus asserts a strict division between a rule-based strategy, where rules express general grapheme-phoneme correspondences and are allowed to overgeneralise, and an memory-based strategy, where the items stored in memory are whole spelling words. To avoid confusion in terminology, we will use the term *grapheme* strictly in the meaning of 'a letter or a cluster of letters that is realised in the phonological transcription as a single phoneme'. Examples of English graphemes are <b>, <ll>, <ea>, and <ough>. When referring to subword units other than graphemes, we will use the term *spelling unit*. It is assumed that words with regular pronunciations can be correctly converted from text to speech via the rule-based route as well as via the lexical route, whereas irregular words can only be converted to speech correctly via the lexical route.

Single-route theory (see, e.g., Glushko, 1979; Seidenberg & McClelland, 1989; Plaut *et al.*, 1996) states that skilled reading aloud is accomplished by a single mechanism, in which a less strict distinction exists between different spelling-phonology correspondence levels. Glushko (1979) introduces the idea of a similarity matching component, that converts spelling strings to their pronunciation by matching them with stored lexical items. The connectionist feed-forward model of Plaut, McClelland, Seidenberg, and Patterson (1996), which is inspired and based on a similar model proposed by Seidenberg and McClelland (1989), is a primary example of a single-route system successfully implemented and able to learn. Plaut *et al.* (1996) train a multi-layered feed-forward network on grapheme-to-phoneme conversion, resulting in a system of which the processing is neither purely lexical nor purely rule-based. Rather, knowledge of text-to-speech correspondences is stored in the automatically learned weights of the connections between the network units. This *distributed* knowledge representation is opaque, in the sense that it cannot easily be interpreted or translated into symbolic terms.

Plaut *et al.*'s (1996) model learns automatically, which allows for training it on other languages than English. The performance results reported by Plaut *et al.* (1996) on test words are in any case quite good and are argued to correlate with naming latencies of both regular and exception words.

The general idea that both the dual-route and the single-route approaches subscribe to, is that a psychologically realistic model of skilled reading aloud has to capture knowledge of correspondences between spelling and speech ranging from the level of whole words to the level of grapheme-phoneme correspondence rules. Lexical reading in isolation would prevent the pronunciation of nonwords, which is clearly possible for human readers. Rule-based reading in isolation (where rules apply to small spelling entities, such as graphemes) does not explain that various experimental findings point to significant word similarity effects in naming latencies (e.g., Glushko, 1979; Jared *et al.*, 1990; Van Orden *et al.*, 1990). Two fundamental questions are whether the proposal by Coltheart (1978) is too restrictive, and whether the proposal by Plaut *et al.* (1996) is too unrestrictive towards the linguistic entities they presume. Coltheart's (1978) dual-route model does not allow for larger word chunks than graphemes to play a role in processing, whereas there are indications that speakers use word chunks such as syllable bodies (Jared *et al.*, 1990; Patterson & Morton, 1985), syllable rimes, or whole syllables (Norris, 1993) as processing units. Syllables, syllable rimes and syllable bodies appear to have a higher psycholinguistic validity in reading aloud than letters, graphemes and phonemes (Liberman *et al.*, 1974; Mehler *et al.*, 1981). Furthermore, there are large differences in sensitivity to spelling-to-phonology units between speakers of different languages (Cutler *et al.*, 1986). Models in which all of the processing units are used that are especially significant and relevant to word naming in the language the model is tuned to, are referred to as *multiple-level* models (e.g., Patterson & Morton, 1985). Of course, such models easily run the risk of being too unrestricted, by allowing any correspondence level to add to the spelling-to-phonology conversion process, a situation that might be opaquely appearing in Plaut *et al.*'s (1996) model.

Norris (1993) describes a computational model that combines a large number of advantageous model features from both dual-route and single-route models, and avoids most of their disadvantages. Norris' model is an connectionist interactive activation network, which performs as a single route system (i.e., there is no internal modularisation). At the input side of the model, several different types of spelling units are presented simultaneously, ranging from letters, via syllable rimes and bodies to whole words. The input thus consists of all elements that might be of importance in the spelling-to-phonology conversion. All information is allowed to interact via excitatory and inhibitory connections. Norris (1993) presents comparative data between his model and human performance, and shows that the correlation is high. Besides, he argues that his model is an example of combining the dual-route and single-route approaches, and can be seen as an implementation of either of them. Again, the question could be asked whether the model is too unrestrictive by allowing so many, possibly unimportant, spelling input units. Furthermore, processing within the model is dependent on a rather large number of parameters (9), and finally, the model does not learn.

In conclusion, one could state that there has been a shift in modelling developments from the very strict classical dual-route model (Coltheart, 1978) and the vague notion of a single-route model (Glushko, 1979), towards the consensus view that a model of skilled reading aloud has to capture spelling-phonology correspondences ranging from graphemes to whole words (see also Humphreys & Evett, 1985; Norris, 1993). The simultaneously growing interest in connectionist modelling of psychological processes has played a major part in this shift. The Seidenberg and McClelland (1989) model, and more recently, the Plaut *et al.* (1996) model have been crucial here. Connectionist learning and knowledge representation is also proposed in Coltheart *et al.*'s (1993) implementation of the dual-route model. Although the specific reasons for adopting connectionist modelling techniques are not the same for every model proposal (actually, none of the models mentioned here incorporate all desirable properties), connectionist modelling is used because (i) it enables learning; (ii) it enables interaction of knowledge at different levels of spelling-phonology correspondences, modelling multiple-level processing (Plaut *et al.*, 1996, as well as Norris, 1993, argue that this is the key factor for the apparent success of their models in accounting human reaction time performance); (iii) it introduces typically connectionist properties that positively discriminate such models from other classical approaches, such as the use of nonsymbolic, distributed knowledge representation, and the consequent resistance to fair amounts of noise or damage.

Although these three advantages are used to put forward connectionist modelling as essential to modelling skilled reading aloud, it remains to be investigated whether there might be other modelling techniques that incorporate these advantages. In fact, within the Artificial Intelligence subdomain of machine learning, several examples can be found of learning algorithms that are able to learn classification tasks on the basis of noisy data. In this paper, we investigate the application of a simple data-oriented, inductive machine learning algorithm to the task of converting spelling words to phonemic transcriptions, and show that there are at least grounds to believe that this technique could in fact incorporate the three advantages of connectionist modelling as well, using relatively simple machinery. The resulting model is in effect a single-route model of skilled reading aloud; we will, however, not focus on the *psychological validity* of the model here nor its relations to models of the *acquisition* of the young reader's skills, but rather on the kind of *knowledge representation* automatically constructed by the model and the way in which it radically differs from the acclaimed nonsymbolic, distributed connectionist representations.

2 SPC: a symbolic, single-route, multiple-level model

2.1 Subword-phoneme correspondences

In this contribution, a non-connectionist single-route model of spelling-to-phonology conversion is presented that uses subword letter strings of variable length as representation units. These subword chunks are not predefined or derived from linguistic morpho-phonological theory, but are automatically extracted from a corpus of word-pronunciation pairs. For a technical description of the learning and processing algorithms involved in the model, the reader is referred to Appendices A-C, containing formulas and pseudo-code descriptions of the algorithms; see also Daelemans and Van den Bosch (1992), and Daelemans *et al.* (1996). Here, we present a more intuitive description of the model by relating it to notions generally used in psycholinguistics.

The model, called the Subword-Phoneme Correspondence (SPC) model, is a single-route model which finds its inspiration from the general notion of *analogy* in the overall correspondence of a writing system and the pronunciation of a certain language, viz. words that are spelled similarly, are pronounced similarly. The SPC model automatically learns to find those parts of words, or subword chunks, on which similarity matching can safely be performed. This is perhaps best illustrated in the example of the pronunciation of the English word <behave>. An 'analogical' model which operates on a certain similarity metric, and which has already encountered (learned) roughly similar words such as <shave>, <beehive>, and <have>, and perhaps even <behave> itself, will certainly have a number of clues as to how <behave> is pronounced. However, in this example, problems may arise with the <a> of <behave>. If the similarity matcher of the analogical model decides to retrieve the pronunciation of the word <have> as the pronunciation of <-have> in <behave>, the incorrect pronunciation /bihæv/ would result. The SPC model does not take such overgeneralisation risks. The model is extremely sensitive to context, in the sense that it will have stored the knowledge that <have> is not enough context to be certain of the pronunciation of the <a>. The SPC model will look for more contextual information. In a present implementation of the SPC model trained on English words, the SPC model decides to take /eɪ/ as output only when it finds the subword chunk <ehave> in the input word. Note that this model has encountered the word <behave> during learning, but that for the case of the pronunciation of the <a>, it was not necessary to store the complete word, as there were no other words with the subword chunk <ehave> with a different pronunciation of the <a>. In sum, the SPC model stores single letter-phoneme correspondences with the context that is sufficient to be certain that the mapping is unambiguous (in the learning material).

Each of these subword-phoneme correspondences can be seen as a context-sensitive rewrite rule, which rewrites a letter to a phoneme. As the context may be of any width, many of these rewrite rules are much more specific than a typical rewrite rule would be; many even contain whole words. Although one would be tempted to categorise the SPC model as a rule-based model, it is just as well possible to view the SPC model as a lexical model. Basically, the SPC model is just a compressed version of the word-pronunciation corpus it is trained on. After training, it contains in a compressed format complete knowledge of the pronunciation of all words of the learning material, except for homophones such as <read> (pronounced as /rɪd/ or /rɛd/), of which only one pronunciation is stored.

To illustrate the appearance of automatically extracted subword-phoneme correspondences (SPCs), Table 1 lists some example SPCs that the model extracted from English data and French data.

From Table 1, it can clearly be seen that some SPCs express very general pronunciation knowledge (e.g., the French <ç> is always pronounced /s/), whereas other SPCs are used to disambiguate between only a few words, e.g., <esiden> – /ɪ/ discriminates <president>

English SPC											Phoneme	Example Word
.	.	.	.	.	v	o	.	.	.	.	v	voucher
.	.	.	e	s	i	d	e	n	.	.	ɪ	president
.	.	.	.	w	o	_	.	.	.	.	u	two
.	.	.	_	h	a	v	e	_	.	.	æ	have
French SPC											**Phoneme**	**Example Word**
.	.	.	.	.	ç	.	.	.	.	.	s	français
.	.	_	_	b	e	a	u	x	.	.	o	beaux
.	.	.	.	v	i	n	_	.	.	.	E	vin
.	.	.	.	n	c	o	.	.	.	.	k	francophone

Table 1: Examples of automatically extracted subword-phoneme correspondences (SPCs), with their associated phonemes, from English and French data. Example words containing the SPCs are given. Dots represent unused context positions; underscores represent word boundaries.

from <reside>.

2.2 Compressing SPC knowledge into a decision tree

The model, which has been implemented and trained on various corpora of the Dutch, French and English language (Van den Bosch & Daelemans, 1993; Daelemans & Van den Bosch, 1993, 1994; Van den Bosch *et al.*, 1994) does not actually store a large list of context-sensitive rewrite rules. It compresses the information contained in these rules even more by storing them in a decision tree. Each rule is represented as a path in this tree. A path consists of a starting node which represents the target letter that is to be mapped to a phoneme; the consecutive nodes represent the consecutive context letters. The order in which these letters are attached to the path is governed by computing their overall relative importance in disambiguating the mapping. This is done using Information Gain, a computational metric from Information Theory. A description of this metric is given in Appendix B. Computation of the Information Gain of context positions renders a result that is consistent for the three corpora of English, French, and Dutch used. Trivially, the focus letter itself is the most important 'context' letter. The further the context position is removed from the focus letter, the less important that position is for disambiguation, on the average. Furthermore, there is an as yet unexplained difference between right and left context: right context positions are computed to be slightly more important than their respective left context positions. In practice, this leads to an ordering in which the first character on the right is the first context expansion, i.e., the first node down the tree. Then follows the first character on the left, then the second character on the right, then the second character on the left, and this alternating pattern simply repeats. To visualise the way in which knowledge is organised in the decision tree, Figure 1 displays the part of the tree in which the pronunciation of the <a> in the word <behave> is stored.

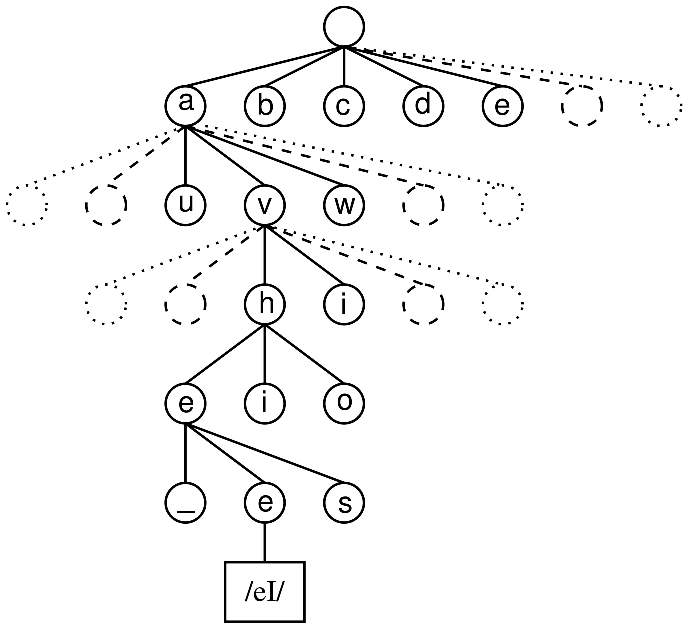

Figure 1: Retrieval of the pronunciation of $< a >$, /eɪ/, of the word $< behave >$. The path represents the minimally disambiguating context $< ehave >$.

With the <a>-node as starting point, the node labelled with the first character on the right, <v>, is the second node accessed in the path. Then, the <h>-node, the first character to the left of the <a>, is taken. At that point, the only possible extensions stored in the tree are <have>, <havi> (from <having>) and <havo> (from <havoc>); the pronunciation at that point is still ambiguous. Then, the <e>-node is accessed, which leaves open the extensions <_have> (underscores depict word boundaries), <ehave>, and <shave>. As mentioned earlier, at the next step, the model, retrieves the unambiguous phonemic mapping /eɪ/, when the final <e> node is reached.

It can be seen that the depth of a path reflects in a certain sense the ambiguity of the mapping it represents. End nodes near the top of the decision tree typically belong to highly regular pronunciations. For example, the French model contains at the top layer of the tree the end node <ç>, as this special character always maps to /s/, regardless of the context. An example of an extremely ambiguous mapping is that of the first <o> of <photograph>, /əʊ/, as opposed to the highly similar <photography>, in which the <o> maps to /ə/. In this case, a context to the right of eight letters is needed to disambiguate between the two pronunciations.

2.3 A best guess strategy when exact matching fails

In the SPC model, all spelling-to-phonology knowledge contained within the learning material is stored lossless, with the exception of homophones, of which only one pronunciation is kept. The rule-based aspect of the decision tree, however, enables the model also to generalise to new cases. To retrieve the pronunciation of a word or nonword[1] that was not in the learning material, each letter of the new word is taken as a starting point of a tree search. The search then traverses the tree, up to the point where the search successfully meets an end node, or where the search fails as the specific context of the new word was not encountered in the learning material, and consequently was not stored as a path in the tree. In the first case, the phonemic label of the end node is simply taken as the phonemic mapping of the new word's letter. In the second case, the exact matching strategy is taken over by a *best guess* strategy.

In present implementations of the SPC model, the best guess strategy is implemented in a straightforward way. When building a path in the tree, the construction algorithm constantly has to check whether an unambiguous phonemic mapping has been reached. At each node, the algorithm searches in the learning material for all phonemic mappings of the path at that point of extension. In cases when there is more than one possible phonemic mapping, the algorithm computes what is the most *probable* mapping at that point. Computation is based on occurrences: the most frequent mapping in the learning material is preferred (in case of ties, a random choice is made). This extra information is stored with each non-ending node. When a search fails, the SPC model returns the most probable phonemic mapping stored in the node at which the search fails.

A pseudo-code description of the decision-tree-construction algorithm, the decision-tree-retrieval algorithm and the best-guess addition described here are given in Appendix C.

2.4 Preparing a corpus: automatic alignment

The SPC model processes words letter by letter. For each letter, it attempts to find a phonemic mapping. However, in many writing systems, such as those of English, French, and Dutch, words exist of which the phonemic transcription contains less phonemes than the word has letters, for example, the English word <book> - /buk/. This is because these words contain graphemes of more than one letter (the <oo> in <book>). The problem is how to present the word to the SPC construction algorithm while avoiding the counter-intuitive mapping of the third letter <o> to the third phoneme /k/. This is done by *aligning* the phonemic transcription to the spelling word, and inserting *phonemic nulls* in the phonemic transcription at those points where in the spelling there is a grapheme containing more than one letter. In the example of <book>, an aligned phonemic transcription of /bu-k/ (the hyphen depicts the phonemic null) would be appropriate. With this alignment, the SPC model has to store the knowledge that only one of the two <o>'s of the word <book> maps to the phonemic mapping /u/, and that the other <o> maps to a phonemic null, which is not realised in the resulting phonemic transcription.

The problem of how to align a phonemic transcription to a spelling word is not trivial. Sejnowski and Rosenberg (1987) use a pre-aligned text-to-speech corpus to train their connectionist NETtalk model, but are not very explicit about the way in which the corpus

[1] Note that from the viewpoint of the SPC model, all unencountered words can be regarded as nonwords, as the model, like most other models, does not take lexical-semantic knowledge into account.

is prepared. In fact, the problem of aligning spelling strings to phonemic strings involves a non-trivial amount of linguistic knowledge engineering, and removes a considerable amount of complexity from the text-to-phoneme conversion problem. Daelemans and Van den Bosch (1994) present a solution to this knowledge acquisition problem by using a simple automatic, data-oriented algorithm that searches for the most probable spelling-phoneme alignment of a word. The model uses a probabilistic association matrix between letters and phonemes, which is based on the target corpus itself. This method results in aligned corpora of which some alignments are counter-intuitive, but not unreasonable. It is important to note that for the SPC model construction, it is not essential that the alignment is (intuitively) optimal; bad or strange alignments only lead to more nodes in the tree, as the corpus becomes less redundant. The strange alignments proposed by the automatic alignment algorithm do not lead to significantly worse performance (Daelemans & Van den Bosch, 1994). In Appendix A, a pseudo-code description of the basis of the automatic alignment algorithm is given.

3 Analysing orthographic depth

Being based on data only, rather than on linguistically motivated rules, the SPC model directly and objectively reflects after construction some of the statistical properties of the data. Trained on different languages, the SPC model behaves very differently, for example, in terms of model magnitude and generalisation performance on unseen words. These comparative data, which basically reflect complexity differences of the learning material, can be used in view of the psycholinguistic concept of *orthographic depth.* This term is used in psycholinguistics to denote the complexity of the relation between a language's writing system and its pronunciation (e.g., Katz and Frost, 1992). On the complex end of the orthographic depth continuum, one finds languages such as Hebrew, of which the spelling of a word presents few clues as to how the word is pronounced. Such writing systems are said to have a *deep* orthography. On the other end of the continuum one finds languages such as Serbo-Croatian, of which the conversion from spelling to phonology is perfectly regular, which are said to have a *shallow* orthography. Frost, Katz & Bentin (1987) report on comparative experiments with naming and lexical decision tasks between speakers of Hebrew, English and Serbo-Croatian, and find that speakers of the shallow language (Serbo-Croatian) appear to use much more a kind of rule-based spelling-to-phonology conversion than do speakers of the languages with the deeper orthographies (English and Hebrew). Seen from the viewpoint of the dual-route model, it is argued that the shallower the orthography of a language is, the more speakers tend to use the rule-based route, simply because this route is very reliable in a regular spelling-to-phonology system. The deeper the orthography, the less use is being made of the rule-based route, although it is never totally unused (Carello *et al.*, 1992). Katz and Frost (1992) refer to this apparent dependency of the ratio between the two routes and the depth of the orthography as the *Orthographic Depth Hypothesis* (ODH).

In Van den Bosch *et al.* (1994), a detailed analysis is given of the application of the SPC

Language	Generalisation Accuracy on Letter-Phoneme Mappings
American English	91.0
French	98.3
Dutch	97.6

Table 2: Generalisation accuracy performance of three SPC models trained on comparable corpora of English, Dutch and French, on correct letter-phoneme mappings.

model to three comparable corpora of English, French and Dutch. For each language, a 20,000-word corpus was constructed containing randomly extracted dictionary words, i.e., the words in the corpora can be of any size, and of any syllabic or morphological complexity. The English and Dutch data were extracted from the CELEX (Burnage, 1990) lexical data bases; these data bases contain more than 20,000 words, but were downsized by random sampling to match the size of the French word corpus, which was extracted from the BRULEX lexical data base (Content *et al.*, 1990). Each word corpus was partitioned once in a 1/13 test set (7.7% of the material) and a 12/13 training set. Although this is an arbitrary partitioning, our primary goal was to measure generalisation performance on unseen material, represented by the test words which are not shown to the learning algorithms during the learning phase. Please note that each test set thus contains randomly picked dictionary words; the test set does not contain especially selected words which display only general or irregular text-to-phoneme mappings, or which are only monosyllabic or monomorphemic. Furthermore, it should be noted that to the learning algorithms, the test words are what nonwords are to experienced human readers; they are similar to existing words (the words in the training set), but are never seen during the training phase, and chances are that many of these words contain unique, unseen letter-phoneme mappings.

When the decision tree is constructed for each language, large differences become apparent in the number of tree nodes: the English model contains 49,064 nodes, the Dutch model contains 24,266 nodes, and the French model contains 16,313 nodes. The number of nodes is a rough indication of the ambiguity of the corpus. The English corpus is the most irregular in the mapping between spelling and (pre-aligned) phonology. The French corpus is the most regular. The differences between the three languages become even more apparent in Figure 2. In this Figure, the number of end nodes (i.e., nodes containing an unambiguous phonemic mapping) per tree depth, or path length, is represented by bars. For example, the largest white bar in the front row, labelled '1–1–2', indicates that, in the French model, most paths end with one left context character and two right context characters.

The generalisation performance on unseen words, as reported in Van den Bosch *et al.* (1994), is listed in Table 2. Generalisation performance on unseen words (for each SPC model, approximately 1,500 words were used for this purpose) is expressed in the percentage of correctly converted letter-to-phoneme conversions of these unseen words.

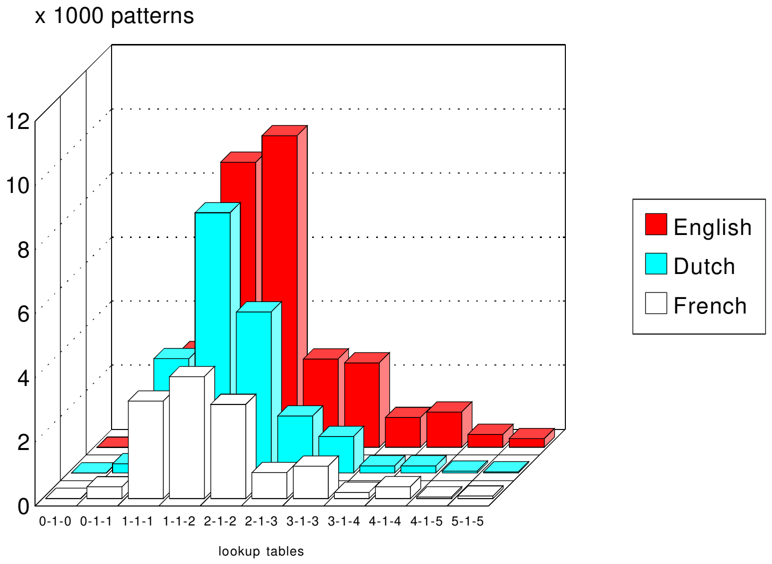

Figure 2: Numbers of end nodes, represented by bars, per tree depth (path length), for three SPC models trained on comparable corpora of English, Dutch and French.

As can be seen from Table 2, the smallest model, i.e., the French model, generalises best to unseen data. The Dutch model performs slightly worse; the English model performs significantly worse than the two other models. One could extrapolate these results by asserting a one-dimensional scale of orthographic depth, where these generalisation performance values mark the ranking of a language (or rather, of a corpus) on this scale. However, as was put forward earlier, the problem of converting spelling to phonology also involves the parsing of words into graphemes, i.e., the spelling-phonology alignment problem. Van den Bosch *et al.* (1994) use the data-oriented alignment algorithm described earlier, trained to align phonemic transcriptions to spelling strings. They find that the French corpus is the most complex corpus to be aligned, and the English corpus the easiest one; the Dutch corpus is slightly more complex than the English. This result is reverse to that obtained with the SPC model. It is mainly caused by the fact that French graphemes are on the average longer and internally more confusible (e.g., <eux>, <aux>, <eau>, <au>) than English or Dutch graphemes. Van den Bosch *et al.* (1994) conclude by proposing a two-dimensional complexity measure denoting the orthographic depth of a corpus, of which the coordinates are the performance results obtained with spelling-phonology alignment, and spelling-to-phonology conversion with the SPC model.

4 Discussion

4.1 Requirements for psychological validity

With a brief overview of the developments within the field of psycholinguistic modelling of reading aloud, we have tried to identify the state of affairs both on a theoretical level as well as on the level of modelling. There has been a large influence from developments with connectionist modelling techniques on the latter. Connectionism offers the tools for learning, for combining several processes into one single-processing system, and for generating robust models that can generalise to new data. The success of connectionist models of reading aloud (Plaut *et al.*, 1996; Seidenberg & McClelland, 1989; Norris, 1993) has even put forward the connectionist paradigm as a theoretical basis, a view that notably Seidenberg and McClelland (1989) have pioneered. Yet, these models have some practical deficiencies, such as the large number of parameters that must be set correctly in order to let Norris' (1993) model perform like human beings do. As far as learning is concerned, there are alternatives to connectionist modelling that lead to automatically learned models which are able to generalise to new unseen words, and which use a sort of knowledge representation that is neither strictly rule-based nor strictly lexical. We have presented such a model. However, this SPC model has not yet been reviewed in terms of the three general requirements for the validity of a psychological model of reading aloud:

(i) A model of reading aloud should be able to learn its knowledge automatically. This statement is still quite distant to the question how knowledge of text-to-speech is acquired by a child; it is far from clear how to relate the learning of a connectionist model, or the SPC model for that matter, to human learning of reading aloud. Sejnowski and Rosenberg (1987) carefully argue that there is a correlation between the learning curves of their NETtalk model and human learning curves, both on a global scale as well as on the level of specific letters, such as the English <c> (Sejnowski & Rosenberg, 1987, pp. 155-156). Not addressing that discussion here, we can safely argue that the SPC model is able to learn its knowledge automatically, without being bound to linguistic engineering or hand-wiring. The storage of knowledge in a tree and the use of a best guess strategy can be further specified as a kind of *lazy learning* (cf. Daelemans, 1995), with which learning is relatively straightforward storage of examples, and performance is some kind of intelligent, example-based similarity matching. For a discussion on the psychological validity of lazy learning, see for example Smith and Medin (1981); Derwing and Skousen (1989).

(ii) The model should have the ability to convert spelling to phonology via multiple levels, using those spelling units that are most relevant, to approach human performance. We have shown that the construction of the SPC model involves a process in which relevant pieces of spelling strings are automatically extracted from example words. The resulting model does not manipulate clear-cut syllable bodies or rimes, but operates on SPCs which can range from very general to extremely specific. The explanatory value (or 'readability') of an SPC model in this respect is not very high, although one could argue that the majority of the SPCs in most SPC models have a letter span of only four letters wide (see Figure 2), which is slightly larger than the average syllable (which is roughly three to four letters wide, averaging over languages); there might be a correlation in this fact. More generally, we would like to argue that the kind of processing in the SPC model is a form of multiple-level processing, that is not even bounded to a number of levels.

(iii) The model should be robust, and should be able to generalise to new data. The SPC model uses a kind of knowledge representation that is symbolic, but also distributed. Converting one spelling word into its phonemic transcription involves many SPCs (i.e., paths in the decision tree); one SPC stores one bit of spelling-to-phonology knowledge that in principle applies to many words. The decision tree stores knowledge that goes from general (small, interpretable rules) to specific (whole-word information), and is shown to be successfully applicable to new words containing partially new letter-phoneme mappings (Van den Bosch and Daelemans, 1993, Van den Bosch *et al.*, 1994).

As regards robustness of the model, cutting the tree at certain points does not lead to a system breakdown. Rather, pruning the tree is a simple yet interesting experimental tool to simulate certain text-to-speech conversion deficiencies. Cutting all paths at a certain depth, for example, would leave the general spelling-to-phonology intact, and would lose the more specific knowledge. It can be imagined that model performance would be similar to that of patients suffering from Surface Dyslexia (Marshall & Newcombe, 1973). Future research should certainly address in this respect the six questions formulated in Coltheart *et al.* (1993) that models of reading aloud should address. Especially the questions concerning Surface, Phonological and Developmental Dyslexia.

4.2 Word pronunciation consistency metrics: a potential contribution of SPC

We have presented SPC, a data-oriented model of the conversion of text to phonemic speech. We have argued that it presents an alternative to connectionist modelling, displaying most of the properties that typically are attributed to connectionist models. Furthermore, we have shown the model to be an indicator of the complexity of grapheme-to-phoneme conversion for a certain corpus. Thus far, we have not analysed the SPC model on its ability to predict the complexity of the phonemisation of single words, nor the relation this metric might have with human reaction time latencies during the pronunciation of these words.

Processing within the SPC model is strongly related to the ambiguity of the pronunciation of a word. Within psycholinguistics, the term 'word pronunciation consistency' is used to denote this average regularity of the pronunciation of a word. There is a tendency within psycholinguistics to focus on syllable nuclei or syllable bodies to determine whether the pronunciation of a word is consistent or inconsistent. This at least correlates with the fact that in general, vowels (i.e., syllable nuclei) are much more ambiguous than consonants as regards their pronunciation.

The SPC model provides an unbiased measure which does not focus on any predefined entity, but which computes for a word simply the sum of the path lengths involved in

Letter in Context											Phoneme	Depth
.	.	.	.	_	f	o	.	.	.	.	f	2
.	.	.	.	f	o	r	k	.	.	.	ɜ:	3
.	.	.	f	o	r	k	_	.	.	.	-	4
.	.	.	.	r	k	_	.	.	.	.	k	2
											Total Depth Sum	11
											Average Depth	2.8
.	.	_	_	_	h	a	v	e	n	.	h	7
.	.	_	_	h	a	v	e	n	_	.	eɪ	7
.	.	.	.	a	v	e	n	.	.	.	v	3
.	.	h	a	v	e	n	_	_	.	.	-	6
.	h	a	v	e	n	_	_	_	_	.	n	8
											Total Depth Sum	31
											Average Depth	6.2

Table 3: Example computation of the average tree search depths (context widths) of retrieved paths for the English words <fork> and <haven>.

finding the phonemic mappings of all of its letters, or the mean of this sum, by dividing it by the number of letters. An example of the computation of these metrics for two English words is given in Table 3.

A word which involves only shallow decision tree searches, as the word <fork> in Table 3, contains relatively unambiguous letter-phoneme correspondences, and could be said to be 'consistent'; when the processing of a word involves some deep tree searches, as with the word <haven> in Table 3, the word could be regarded 'inconsistent'. It should be noted that this consistency metric is not directly based on the regularity of a word as compared to the pronunciation of similarly spelled words, which might have partly agreeing or disagreeing pronunciations; the SPC model does not operate on the lexical level. Rather, the metric takes into account on a detailed level all *subword disagreements* between the target word and all other words in the learning material.

There are basically two existing alternatives to this word pronunciation consistency metric, viz. neighbour word counts and mean bigram frequency. Neighbour word counts are used under the assumption that the process of pronouncing a word is influenced by numbers and/or frequencies of similar words. Jared *et al.* (1990) proposes a function for estimating the complexity of a word's phonemisation as opposed to neighbour words by counting frequencies of *friends*, i.e., neighbour words with identical phonology, and *enemies*, i.e., neighbour words with conflicting phonology. A significant effect of longer naming latencies was found with words with high frequency enemies, as well as shorter naming latencies with words with high frequency friends (Jared *et al.*, 1990). These experimental results appear to point at a lexical effect of activation among friends, and competition between enemies, at the level of words. This effect is in fact modelled by the DRC implementation of the dual route model (Coltheart *et al.*, 1993).

At the subword level, *mean bigram frequency* is commonly used to measure a word's overall similarity to other words. It does not take into account phonological inconsistency of similar words, nor their frequency. Mean bigram frequency of a word is the average of the occurrences of all of its position-specific bigrams (e.g., <save> contains the bigrams <sa>, <av> and <ve>). Bigram occurrences are computed on the basis of a large corpus. This way, the metric expresses the 'typicality' of a spelling. Experiments with varying mean bigram frequency have rendered no significant naming latency effects (Andrews, 1992).

Since the SPC word metric has only been tested in preliminary word naming experiments, we can only evaluate its possible validity by comparing it on an abstract level to the concepts of frequency of friends and enemies, and mean bigram frequency. In comparison with the latter, the SPC word metric does take phonological consistency into account. In fact, it is a direct measure of letter-phoneme correspondence consistency. In comparison with metrics of friends and enemy words, the SPC model is not able to give direct information on numbers of words, as the knowledge in the SPC model is not word-oriented. If the information is present in the metric, it is only very indirectly. Word frequency information is not available in the SPC model either. Analogous to the findings of Andrews (1992), it might be the case that a naming latency effect could be found using the SPC word metric, but only between consistent low-frequency words versus inconsistent low-frequency words.

4.3 Future research

In this paper, we have not addressed the dependencies of the spelling-to-phonology domain on related morpho-phonological domains such as word morphology and word stress assignment. Morphology has a notoriously noisy influence on the pronunciation of morphologically complex words in some languages (e.g., Dutch, English), as it may overrule grapheme-phoneme regularities. See, e.g., the English example <photograph> vs. <photography> that was given earlier. Word stress assignment is also influenced by morphology, and is at least interrelated with grapheme-phoneme conversion. Any model that is aimed at modelling word pronunciation should incorporate stress assignment; most models mentioned in this paper do not (an exception is NETtalk, Sejnowski & Rosenberg, 1987); their correlation results with human data are mostly obtained with training their models on simple monomorphemic, monosyllabic words (Seidenberg & McClelland, 1989; Plaut *et al.*, 1996; Norris, 1993; Coltheart *et al.*, 1993). Future extensions to the SPC model should include the incorporation of morphology and stress assignment; isolated data-oriented accounts of both domains already exist (e.g., see Daelemans *et al.*, 1994; Gillis *et al.*, 1993; Van den Bosch *et al.*, 1996). Further experimental research should address the question whether the word pronunciation complexity measure described in the previous subsection is psychologically valid.

Acknowledgements

The modelling research reported in this paper was carried out while the first author was affiliated to the Institute for Language Technology and AI at Tilburg University, the Department of Psychology at Tilburg University, and the Laboratoire de Psychologie Expérimentale at the Université Libre de Bruxelles. Thanks are due to Alain Content, Beatrice de Gelder and Jean Vroomen for ideas and discussions, to Jaap van den Herik for valuable comments on the text, and to an anonymous reviewer for helping clarify some of the psychological issues raised in the paper.

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.

Burnage, G. 1990. *CELEX: A Guide for Users.* Centre for Lexical Information, Nijmegen.

Carello, C., Turvey, M., & Lukatela, G. (1992). Can theories of word recognition remain stubbornly nonphonological? In *Haskins Laboratories Status Report on Speech Research* (pp. 193–204). Haskins Laboratories.

Coltheart, M., Davelaar, E., Jonasson, T., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and Performance VI.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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., & Halter, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, 100, 589–608.

Content, A., Mousty, P., and Radeau, M. (1990). Brulex: Une base de données lexicales informatisée pour le français écrit et parlé. *L'Année Psychologique*, 90, 551–566.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in segmentation of French and English. *Journal of Memory and Language*, 25, 385–400.

Daelemans, W., A., Gillis, S., & Durieux, G. (1994). The Acquisition of Stress, a data-oriented approach. In *Computational Linguistics* 20 (3), 421–451.

Daelemans, W. 1995. Memory-based lexical acquisition and processing. In: P. Steffens (Ed.), *Machine Translation and the Lexicon*, Springer Lecture Notes in Artificial Intelligence 898, 85–98.

Daelemans, W. & Van den Bosch, A. (1993). Tabtalk: Reusability in data-oriented grapheme-to-phoneme conversion. In *Proceedings of Eurospeech '93* (pp. 1459–1466).: Berlin: T.U. Berlin.

Daelemans, W. & Van den Bosch, A. (1994). A language-independent, data-oriented architecture for grapheme-to-phoneme conversion. *Proceedings of the ESCA-IEEE conference on Speech Synthesis*, New York, 199–203.

Daelemans, W., Van den Bosch, A., & Weijters, A. (1996). IGTree: using trees for classification in lazy learning algorithms. To appear in *Artificial Intelligence Review.*

Derwing, B. L. & Skousen, R. (1989). Real time morphology: Symbolic rules or analogical networks. *Berkeley Linguistic Society*, 15, 48–62.

Frost, R., Katz, L., & Bentin, S. (1987). Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 104–115.

Gillis, S., Durieux, G., Daelemans, W., & Van den Bosch, A. (1993). Learnability and markedness: Dutch stress assignment. In *Proceedings of the 15th Conference of the Cognitive Science Society 1993, Boulder, CO* 452–457.

Glushko, R. (1979). The organisation and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 647–691.

Heemskerk, J. (1993). A probabilistic context-free grammar for disambiguation in morphological parsing. In *Proceedings of the Sixth Conference of the EACL, Utrecht, The Netherlands*, 183–192.

Humphreys, G. W. & Evett, L. J. (1985). Are there independent lexical and nonlexical routes in word processing? An evaluation of the dual-route theory of reading. *The Behavioural and Brain Sciences*, 8, 689–710.

Jared, D., McRae, K., & Seidenberg, M. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, 29, 687–715.

Katz, L. & Frost, R. (1992). The reading process is different for different orthographies: The orthographic depth hypothesis. In *Haskins Laboratories Status Report on Speech Research 1992* (pp. 147–160). Haskins Laboratories.

Liberman, I., Schankweiler, D., Fisher, D., & Carter, D. (1974). Reading and the awareness of linguistic segments. *Journal of Experimental Child Psychology*, 18, 202–212.

Marshall, J. & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research*, 2, 175–199.

Mehler, J., Dommergues, J., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behaviour*, 20, 298–305.

Norris, D. (1993). *A quantitative model of reading aloud.* Technical report, MRC Applied Psychology Unit, Cambridge, UK.

Nunn, A. & van Heuven, V. J. (1993). Morphon, lexicon-based text-to-phoneme conversion and phonological rules. In V. J. van Heuven & L. C. Pols (Eds.), *Analysis and Synthesis of Speech: Strategic Research Towards High-Quality Text-to-Speech Generation.* Berlin: Mouton de Gruyter.

Patterson, K. & Morton, J. (1985). From orthography to phonology: An attempt at an old interpretation. In K. Patterson, J. Marshall, & M. Coltheart (Eds.), *Surface Dyslexia: Neuropsychological and Cognitive Studies of Phonological Reading* (pp. 15–34). London: Erlbaum.

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

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

Sejnowski, T. J. & Rosenberg, C. S. (1987). Parallel networks that learn to pronounce english text. *Complex Systems*, 1, 145–168.

Smith, E. & Medin, D. (1981). *Categories and Concepts.* Cambridge, MA: Harvard University Press.

Van den Bosch, A. & Daelemans, W. (1993). Data-oriented methods for grapheme-to-phoneme conversion. In *Proceedings of the 6th Conference of the EACL* (pp. 45–53).

Van den Bosch, A., Content, A., Daelemans, W., & De Gelder, B. (1994). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics*, 1, 178–188.

Van den Bosch, A., Daelemans, W., and Weijters, A. (1996). Morphological Analysis as Classification: an Inductive-Learning Approach. To appear in *Proceedings of New Methods in Language Processing, NeMLaP-2*, Bilkent University, Turkey.

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

Appendices

A. Automatic Alignment

Given a corpus of unaligned pairs of words and their phonemic transcriptions, an association matrix is built between spelling letters and phonemes. In pseudo code:

Procedure **COMPUTE-ASSOCIATION-SCORES**

Input: A corpus C of unaligned word-transcription pairs

Output: An association matrix between all letters and phonemes in C

1. Initialise the association matrix M between all letters l in C and all phonemes p in C, to $M_{lp} = 0$
2. For each word W and its associated phonemic transcription P,
   (a) Let $d = |W| - |P|$ (i.e., d is the length difference between W and P)
   (b) For each possible alignment of W and P, do
   For $x = 1$ to $|W|$ do
      - $M_{W_x,P_x} = M_{W_x,P_x} + 8$
      - $M_{W_x,P_{x-1}} = M_{W_x,P_{x-1}} + 4$
      - $M_{W_x,P_{x-2}} = M_{W_x,P_{x-2}} + 2$
      - $M_{W_x,P_{x-3}} = M_{W_x,P_{x-3}} + 1$

The association scores thus computed are converted into probabilities, and the most probable alignment for each word-pronunciation pair is computed.

B. Information Gain

The main idea of *information gain weighting* is to interpret the training material as an information source capable of generating a number of messages (the different phoneme labels) with a certain probability. The information entropy of such an information source can be compared in turn for each feature to the average information entropy of the information source when the value of that feature is known (features in this case are context positions).

Database information entropy is equal to the number of bits of information needed to know the category given a pattern. It is computed by equation (1), where p_i (the probability of category i) is estimated by its relative frequency in the training set.

$$H(D) = -\sum_i p_i log_2 p_i \tag{1}$$

For each feature, it is now computed what the information gain is of knowing its value. To do this, we compute the average information entropy for this feature and subtract it from the information entropy of the database. To compute the average information entropy for a feature (equation 2), we take the average information entropy of the database restricted to each possible value for the feature. The expression $D_{[f=v]}$ refers to those patterns in the database that have value v for feature f, V is the set of possible values for feature f. Finally, $|D|$ is the number of patterns in a (sub)database.

$$H(D_{[f]}) = \sum_{v_i \in V} H(D_{[f=v_i]}) \frac{|D_{[f=v_i]}|}{|D|} \tag{2}$$

Information gain is then obtained by equation (3).

$$G(f) = H(D) - H(D_{[f]}) \tag{3}$$

C. Decision Tree

C.1. Decision tree construction

On the basis of a corpus of aligned word-pronunciation instances (i.e., fixed-length windows of letters associated with phonemes), and given the information gain values of all letter positions computer over the full corpus, a decision tree is built in which instances are stored in a compressed fashion (i.e., as partially overlapping paths of variable length). In pseudo code:

Procedure **BUILD-IG-TREE**:

Input:

- A training set T of instances (letter windows) with their associated phonemes (start value: a full instance base),
- an information-gain-ordered list of features (tests) $F_i ... F_n$ (start value: $F_1 ... F_n$).

Output: A subtree.

1. If T is unambiguous (all instances in T map to the same phoneme p), or $i = (n+1)$, create a leaf node with unique phoneme label p.
2. Otherwise, until $i = n$ (the number of features)
    - Select the first feature (test) F_i in $F_i ... F_n$, and construct a new node N for feature F_i, and as default phoneme p (the phoneme occurring most frequently in T).
    - Partition T into subsets $T_1 ... T_m$ according to the values (letters) $v_1 ... v_m$ which occur for F_i in T (instances with the same letters for this feature in the same subset).
    - For each $j \epsilon \{1, ..., m\}$:
      if not all instances in T_j map to phoneme p, BUILD-IG-TREE (T_j, $F_{i+1} ... F_n$), connect the root of this subtree to N and label the arc with letter value v_j.

C.2. Decision tree processing and retrieval

After tree construction, the phonemic transcription of a word can be looked up in the tree. The word is converted by windowing into fixed-length instances. Each instance is matched against paths in the decision tree, leading to a leaf node with a unique phoneme label or, if no unambiguous mapping can be found, to the probable mapping at the point of tree search failure. In pseudo-code:

Procedure **SEARCH-IG-TREE**:

Input:

- The root node N of an subtree (start value: top node of a complete IGTree),
- an unlabeled instance I (a letter window) with information-gain-ordered feature values (letters) $v_i...v_n$ (start value: $v_1...v_n$).

Output: A phoneme label.

1. If N is a leaf node, output default phoneme p associated with this node.
2. Otherwise, if test F_i of the current node does not originate an arc labeled with letter value v_i, output default phoneme p associated with N.
3. Otherwise,
   - new node M is the end node of the arc originating from N with as label letter v_i.
   - SEARCH-IG-TREE $(M, v_{i+1}...v_n)$