2. Connectionist Models of Language Processing

Connectionist models have been widely used in language processing modelling[1]. For our purposes of modelling bilingual lexical organization and access, the most interesting models are those related to word recognition.

A main feature of connectionist models is the way they represent information. In local representation models a particular feature is represented by a single unit, which acts as a feature-detector that activates when the feature is present. In contrast, distributed representation models represent features by a pattern of activation through a set of nodes (Quinlan, 1991). Both types of scheme have been used for modelling word recognition, especially lexical access. Local representation models use a single node to represent each lexical entry, and the set of these word-nodes forms the lexicon. Distributed representation models of word recognition have a lexical level with a number of nodes. Words are represented at that level by the pattern of activation of the nodes.

The choice for one or the other type of representation is critical for the purposes of this project, because of its implications when representing words of different languages. As will be seen in next pages, a distributed representation seems to be more appropriate than a local one when a single network has to deal with two different sets of words, each set having a different relation between orthography and phonology.

In this section both local and distributed representation models for word recognition are briefly reviewed.

1. See Quinlan (1991) for a review.

2.1. Local Representation Models in Word Recognition

Local Representation models for word recognition are called Word-Detector Models (Monsell, 1991) or Lexical Instance Models (Stone & Van Orden, 1989), in reference to the feature described above. In these models each lexical entry is represented by a single unit.

The first model that appeared in the literature suggesting a parallel activation of the lexical entries is the Logogen Model of Morton (Morton, 1969). In the Logogen model the lexicon is composed by lexical nodes called Logogens. Each logogen is a representation of a word, and it receives activation when this word is perceived either visually or auditorily. Word recognition takes place when the activation of a logogen reaches a pre-determined threshold. This threshold is set according to the frequency of occurrence of the word for the speaker: a very frequent word is represented by a logogen with a low threshold and is recognized faster than a word of lower frequency. The Logogen model was a starting point for recent models of parallel activation.

A model inspired directly on the Logogen model is the Interactive Activation Model of Word Recognition (I.A.) (McClelland and Rumelhart, 1981). Unlike the Logogen model, the I.A. model was developed only for visual word recognition. As figure 2.1 shows, I.A. is a network of three layers with three types of detector nodes. The first layer is composed by nodes sensitive to orthographic features; the second layer is composed by letter detectors; and the third layer corresponds to the level of words. When certain orthographic features are detected in the visual input, some of the first layer nodes are activated. The activation spreads to the letter nodes, and from them to the layer of word nodes. When one of the word nodes reaches the maximum activation, the word it represents will be recognized. This model could account very well for some aspects of word recognition, like frequency and neighbourhood effects (McClelland & Rumelhart, 1981; Seidenberg, 1987).

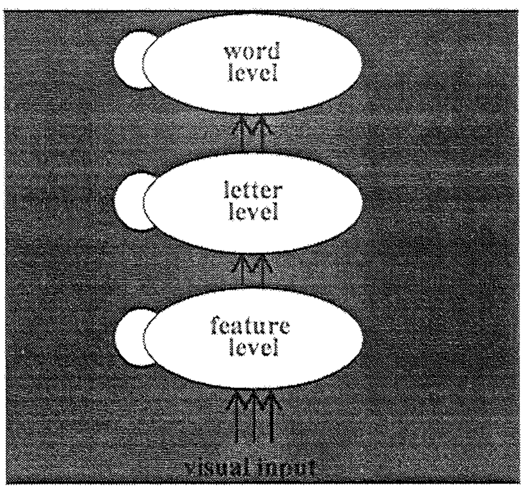

Figure 2.1. The Interactive Activation Model (McClelland & Rumelhart, 1981)

Another model based on the Logogen model is the Activation and Verification Model of Paap and his collaborators (Paap, McDonald, Schvaneveldt & Noel, 1986). The Activation and Verification models incorporate some features of the word-detector models and some from the serial search models. In a first step, a set of word-nodes is activated in parallel, and some of these nodes reach a certain level of activation. Then the Verification process selects among this set of words the best match to the stimulus in a restricted serial search.

Word Detector models have a few limitations that have been pointed out by some researchers. With respect to the Serial Search Models, Word Detector Models account better for lexical effects such as neighbourhood and frequency (Seidenberg, 1987; Monsell, 1991), but they cannot account for other effects. Monsell (1991) emphasized that they do not explain how a new word is acquired, and Seidenberg (1987) pointed out that phonological effects in visual word recognition are not described. These models have been mainly developed for visual word recognition as a process independent of phonology.

The role of phonology during visual word recognition is a polemic issue within the field of word recognition. The Dual Route Model (Coltheart, 1978, and further versions) suggested that there are two word recognition paths, one based on visual information and another based on auditory information, which operate independently of one another. Each of these paths has its own lexicon containing the orthographic or the phonological representation of the words, and word recognition process takes place through one of these routes. Nevertheless, there is important evidence showing the interaction between orthographic and phonological codes during word recognition, indicating that multiple-code activation occurs in both auditory and visual word recognition (Perfetti & Bell, 1991; Seidenberg et al. 1984; Seidenberg, 1985; Seidenberg, 1987; Tannenhaus et al. 1980; Van Orden, 1991; Van Orden, Pennington, & Stone, 1990). Seidenberg (1985) described the main weaknesses of the Dual Route Model using English as an example. He claimed that it is very difficult to define a set of rules for the correspondence between spelling and sound in English; hence a phonological route would often lead to mispronunciations. Another weakness of Dual Route Models is that they define the two paths for visual word perception but do not specify under which conditions one or the other path is selected.

As an alternative to Dual Route Models, Seidenberg *et al.* (1984) proposed the Time Course Model for Word Recognition. This model is based on McClelland & Rumelhart's Interactive Activation Model described above. Taking the original structure of I.A., they incorporated a new set of phonological units attached to the orthographic units. Thus, in the course of visual recognition of the orthographic features, the nodes of the correspondent phonological features are automatically activated.

This model was the first step for the model Seidenberg and McClelland developed in 1989. Their new model has two main features that distinguish it from the models described up until now. In the first place, the model of Seidenberg and McClelland uses a learning rule and their network learns the words during a training phase. The word detector models, as already mentioned, did not describe how the words are acquired: they were implemented by setting the nodes and the weights of the connections between the nodes from the starting point. In contrast, Seidenberg & McClelland's model starts with random weights that are modified during the training phase. The second feature is the use of distributed representations: the words are represented through the activation of the hidden layer of units. A detailed description of this model follows.

2.2. Distributed Representation Models of Word Recognition

Seidenberg and McClelland (1989) elaborated and partially implemented a new model with the goal of accounting for the interaction of phonology and orthography during the process of word recognition and naming. They proposed a new lexical organization that could account for the automatic activation of phonological coding during word recognition. The principal feature of this model is that lexical access representations are not local but distributed. Seidenberg and McClelland's (1989) model is described next.

2.2.1. A Distributed Model for Word Recognition and Naming (Seidenberg & McClelland, 1989)

Seidenberg and McClelland's (1989) model for language processing is depicted in figure 2.2. Their model is a general framework for language processing, but only the part that accounts for lexical access was implemented (framed in the figure).

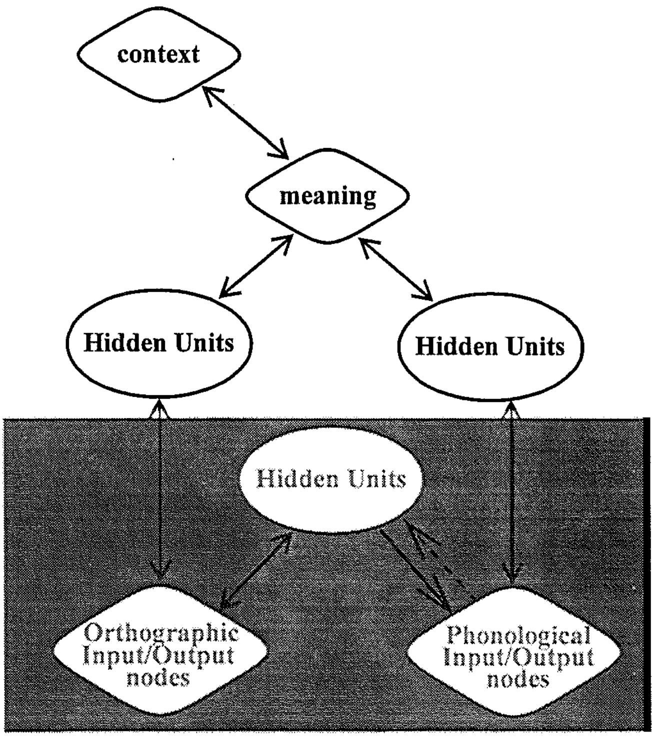

Figure 2. 2. General framework for language processing proposed by Seidenberg & McClelland (1989)

Seidenberg & McClelland designed the network as an Autoencoder, that is, the network gives as output the same pattern received as input. Hence, the number of output units is the same as the number of input units. The hidden units constitute an internal representation of both coding modalities (orthography and phonology).

Results from empirical research seem to indicate that while the processing of phonological information is language-specific (Cutler *et al.*, 1986; 1989;1992), the processing of orthographic information is universal (Seidenberg, 1985; Sebastian-Galles, 1991). The model of Seidenberg & McClelland (1989) provides a framework for considering both universal and language-specific aspects of word processing (cf. Seidenberg & McClelland, 1989, pp. 559).

The results obtained with the simulations confirmed the expectations. The model could process irregular words as well as regular words, pronounce new items, model aspects of reading skills, and simulate human results in tasks such as lexical decision and naming.

The criticism from other researchers (Besner, Twilley, McCann, and Seergobin, 1990) suggested that the performance of the network was not as good as claimed, especially concerning the naming of new words and regarding some effects reported in the literature, as the pseudohomophone effect. Besner *et al.* (1990) were not totally against distributed representations for lexical entries (see op. cit., pp. 445), but they claimed that lexical access requires some routines that the connectionist model of Seidenberg & McClelland did not include. Seidenberg & McClelland (1990) replied that performance problems were due to the small training set (less than 3000 words) and the phonological coding, which was not really suitable for the purposes of the model. They admitted that the model was limited and needed further development, but as discussed in next subsection, their model seems suitable for modeling bilingual lexical organization.

2.2.2. Bilingual Lexical Organization

The Seidenberg & McClelland (1989) model accounts for orthography/ phonology interaction and allows the learning of new words without the need to add more processing units. These two features make the it suitable for modelling bilingual lexical organization and access. Furthermore, if such a bilingual model appears to be successful in dealing with two languages, it will provide support for the distributed representation model of word recognition and suggest further research. For the purposes of this project, only the formal level of the representation of the words is concerned. Thus, only the same part that Seidenberg & McClelland implemented in their model is used in this project. Actually, no major changes have been made, because the original model fulfils several assumptions that will be further discussed.

The main reasons for choosing this model for bilingual modelling have already been mentioned. The first reason is that the distributed representations of the model offer a better explanation for the acquisition of new vocabularies: the words of the new language can be represented using the same number of units, whereas within the local representation framework a new unit is needed for each new word acquired.

The second reason is based on the empirical evidence presented in Section 5.3. The Masked Priming Paradigm in Cross-Language Experiment: The Cognate Effect (Chapter 1), about the differences found in Response Times in form-priming experiments for words of different languages which do or do not share morphological features. A distributed representation that merges orthographic and phonological information can yield the cognate effect: words with similar orthography and phonology are represented in the model by a similar pattern of activation throughout the hidden units, and the change from an internal representation to another just implies little changes in the activation of the units. However, in the case of words with a different form, the pattern of activation of the hidden units is very different, and to switch from the representation of word to the other should take a longer time.

In the next section the model of bilingual lexical organization based on Seidenberg & McClelland's model is described. This new model is called Bilingual Lexical Representations model (BAR).

3. Bilingual Access Representations Model: Version 1 (BAR 1)

The Bilingual Access Representations model (BAR) is a model for bilingual lexical organization which was conceived to model the learning of new words during second language acquisition. As BAR does not have any representation for semantics, it is not a translation model: BAR learns sets of correspondences between orthographic and phonological representations of two different languages. The first Bilingual Access Representations Model (BAR 1) is a pilot model aiming to check the design of the network and the use of different encoding schemes for the input and output units.

This section contains a description of the first simulation carried on with BAR 1 in four subsections. Subsection 3.1. describes the technical features for its implementation; Subsection 3.2. explains the main features for the training phases; Subsection 3.3. shows the results obtained and the analyses performed; and finally Section 3.4. discusses the cognitive validity of the model according to the results.

3.1. Implementation

3.1.1. Learning Algorithm

The learning algorithm used in BAR 1 was the Back Propagation algorithm, as in Seidenberg & McClelland's model. During the training phase, this algorithm allows the connections of the units to adjust their weights by comparing the desired output activation with the actual output of the net.

3.1.2. Network Architecture

BAR 1 had basically the same architecture as was implemented in Seidenberg and McClelland's (1989) Word Recognition and Naming model, as depicted in figure 2.3.

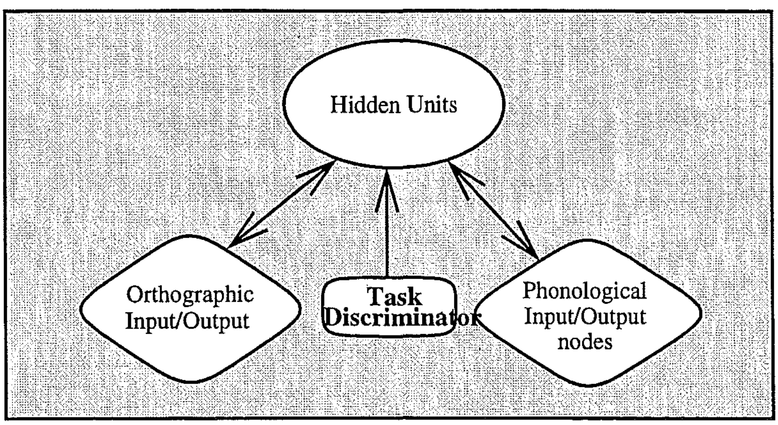

Figure 2. 3. BAR 1 architecture

The difference with respect to the original model of Seidenberg and McClelland is that BAR 1 had to deal with bilingual information. A new set of input units is added in order to allow the network to discriminate the type of input received, in terms of language of the input and number of languages in the training set (one or two). These units were called Task Discriminator[1].

The total number of units of the network was determined by the number of input units required, which was in turn determined by the coding used for input and output. The coding used for BAR differed from Seidenberg and McClelland's (1989) coding. Next subsection gives a detailed description of the encoding scheme used.

3.1.3. Coding

The first encoding used in a pilot version of BAR was the same as Seidenberg & McClelland used in their model, which is fully described in Appendix 1. This encoding did not work in BAR: it resulted in too a large network that could not learn. As Seidenberg & McClelland's coding received some criticism, Seidenberg (1987) suggested that the coding proposed by McWhynney & Leinbach (1991) could be a possible substitute for the original. McWhynney & Leinbach's coding was also adapted for BAR 1 and it is described next.

McWhynney and Leinbach (1991) proposed the new coding scheme in their

1. As will be explained later, the main reason why the Task Discriminator Units are added to the model is to avoid Catastrophic Interference. Nevertheless it should be mentioned that the differences in orthography/phonology interaction between languages would probably allow the network to discriminate which is the language of the input received.

revision of the verb learning model of Rumelhart & McClelland (1980), based on the coding they used in a former simulation (MacWhinney, Leinbach, Taraban, & McDonald, 1989). As this model needed only a phonological input, the coding they proposed was only for phonological coding. They used a set of slot/feature units, where each phonological input unit codes for two types of information: the position of the unit in a syllabic template and the value of some particular distinctive feature. They wanted to emphasize the relevance of syllable structures in language acquisition and learning.

The encoding scheme proposed by McWhynney and Leinbach (1991) consists on two templates, a right-justified and a left-justified template. The left-justified template takes the form: CCCVVCCCVVCCCVVCCC, where C stands for Consonant and V stands for vowel. This pattern codes a full trisyllabic structure in left-to-right fashion. The right-justified template takes the form VVCCC. This pattern only represents the code of the final syllable. The right template is explained by the purpose of their model: the last syllable is very important when learning the past tense, because the morphological mark for the past tense is generally at the end of the word.

The templates, filled with characters and phonemes, were used as input to the neural network designed by Rumelhart and McClelland (1986). To make the templates suitable as input, each character and each phoneme was translated to a certain binary code. For the phonological representation McWhynney and Leinbach (1991) represented each vowel as a combination of 8 distinctive features, and each consonant as a combination of 10 distinctive features. In total, they needed 214 slot/units for the phonological input of the network.

This phonological coding proved to be useful for their simulation of the Verb Learning model. It was able to cope with some of the representational problems in the coding used by Rumelhart and McClelland (1986), and Seidenberg (Daugherty & Seidenberg, 1992) used an adapted version of this coding as well in a new revision of the Verb Learning Model.

The coding described above for the BAR model was partially adapted. Because of its structure, BAR needed both orthographic and phonological information. The scheme of templates proposed by McWhynney and Leinbach (1991) was incorporated to code the orthographic information. This type of coding is very suitable for orthography because the slot/feature code provides information about the position of the characters. The phonological coding was also adapted: in order to keep the number of input units as low as possible, the phonological features were not included. Thus, a unique binary code was given to each character and each phoneme (see Appendix 2 for coding equivalencies).

For the equivalencies of representation, the orthographic coding distinguished 6 possible vowels and 22 possible consonants. Each vowel was represented with 3 units, and each consonant with 5 units. In the left template 12 consonants and 6 vowels are represented; and in the right template 3 consonants and 2 vowels. Thus, the total number of units needed to code the orthographic representation was 99.

The phonological coding distinguished 39 vowels[1] and 32 consonants. For the phonological coding 6 units per vowel and 6 units per consonant were needed. This is equivalent to 138 units in the phonological templates.

In sum, 237 units were needed to code a word. Compared to the original model of Seidenberg & McClelland (1989), the number of input units is very

1. Diphthongs are included in the set of vowels.

reduced.

As already mentioned when describing BAR 1's architecture, in addition to the units needed to code the word, BAR 1 had a set of 3 context input units to indicate the language that was being learnt and which kind of task (see 3.2.3. for further explanation). Thus, the total number of Input units is 240. Except for the context units, the output required was the same as the input. Consequently, the number of Output Units was 237.

3.1.4. Hidden Units

The number of hidden units for BAR was set experimentally, testing different numbers and examining the results after some training, in order to fit two requirements, one concerning the network's implementation and the other the network's performance. The first requirement was that the number of hidden units had to be as low as possible, because in small networks the performance is better and the analysis of the activation is simpler. The second requirement was that high-frequency words should be learnt faster than low-frequency words during the training phase.

After testing different number of hidden units using a training set of Dutch-Dutch pairs of words, the most efficient number of hidden units appeared to be 85. When the number of units was set above 100 the performance of the network was too good, because low-frequency words were learnt as fast as high-frequency words. Additional testing showed that with fewer than 85 the two different phonology/orthography interactions could not be learnt. Figure 2.4 shows the final design of the network with the total number of units needed.

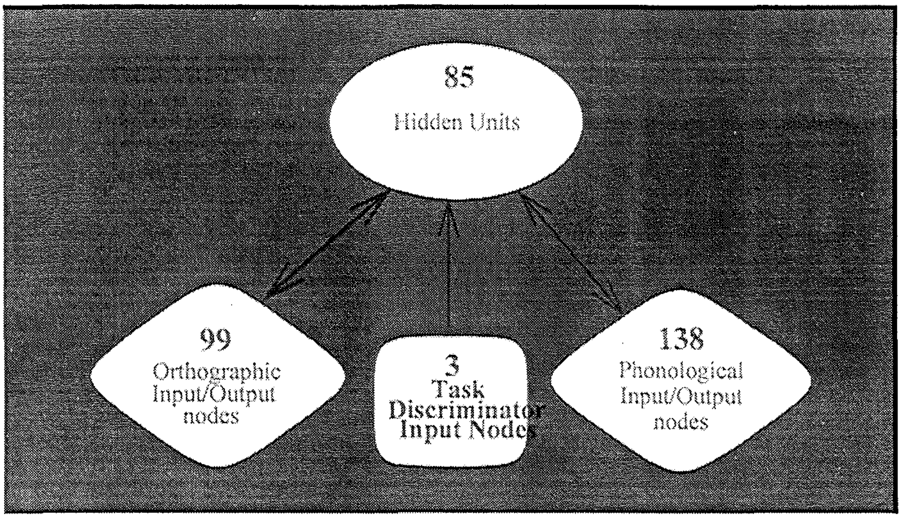

Figure 2. 4. BAR 1architecture and number of units

As the figure shows, 99 units encoded the Orthographic Input and the Orthographic Output, 138 units encoded the Phonological Input and the Phonological Output, and 3 Units were needed for the Task Discriminator Input units. The total numberof units used in BAR 1 is 562.

3.1.5. Parameters

Several parameters were initiated before the simulation began. The learning rate was set at 0.05,and the momentum parameter was set at 0.90, which are the standard values. The network started with small random weights because the system cannot learn if all weights start at equal values. This procedure is called symmetry breaking.

The activation function is a sigmoid function which has a range that lies between 0 and 1. Because it can never reach 0 or 1, the actual target for elements set to 0 in the pattern list became 0.1; and the target for elements set to 1 in the pattern list became 0.9.

3.2. Training

The training of the BAR model had to be designed according to the aim of the simulations. As BAR 1 intended to model human learning of second language words, the training could be designed to model different types of bilingual learning. For example, the network could be trained in two languages from the start in order to model the learning of a person who is in contact with two languages from early childhood. The actual type of training chosen for BAR 1 intended to model the most common case of second language learning, which is the case where a person learns a second language later than his or her mother tongue.

Accordingly, BAR was trained in two phases. During the first phase the network was presented with Dutch words. When the learning rates achieved a certain level of accuracy, the weights of the connections were stored. The second phase consisted in training with English words using the stored weights as starting point.

However, BAR 1 was the first version of the model and our main interest was to examine how it could deal with two different sets of orthographic and phonological information, as well as with two different sets of correspondences between orthography and phonology. The training does not reflect exactly the situation of a person learning a second language, because the size of both training sets of words (Dutch and English words) was comparable.

Next, the different aspects considered in the design of the training phase are described.

3.2.1. Set of Training Words

For both training phases the set of words was obtained from the CELEX database (Burnage, 1990). This database contains, among other features of the lexical entries, detailed information on orthography, phonology and word frequency for the Dutch, English and German languages.

The words for each training set were selected according to their lexical status (all of them were nouns) and their frequency. The selected set of Dutch words was composed of 3738 nouns and the English set was composed of 4249 nouns.

3.2.2. Frequency

As in Seidenberg and McClelland's (1989) model, the frequency of the words was coded in order to provide the network with a training closer to human learning. Thus, the words in each set were presented to the network with a certain probability based on the CELEX word count.

The word frequency within the corpus is scaled down to a range of 1 to 1.000.000 but the range of variation in presentation frequencies is rather high. For example, according to CELEX corpus the word 'one' appears 345 times as frequently as the word 'balloon'. This range was narrowed with a logarithmic frequency transformation. The following formula transforms the estimated frequency of a word to its chance of being presented within an epoch.

$$p = K\log\langle frequency + 2\rangle$$

Seidenberg and McClelland proposed that the value of *K* has to be chosen such that the most frequent word has a probability of about 0.93. In this case the word 'one' is the most frequent word in the training set with a frequency of 2073. A probability of 0.93 for 'one' is obtained by setting the value of *K* to 0.28. Thus the transformation used for the frequency of the words in the training set was:[1]

$$p = (0{,}28)\log\langle frequency + 2\rangle$$

After compressing the range by using this logarithmic frequency transformation, 'one' is presented only about 3.74 times as often as 'balloon'. According to Seidenberg and McClelland (1989) the compression is justified by the fact that the frequency counts in the corpus are based on samples of written and spoken sources coming from adult sources, and that children learning to read and write are not exposed to such a big range. Neither is a person who learns a second language.

1. Although the formula was calculated according to the frequency values of English words, it gave similar probabilities when applied to the frequency values of Dutch words. Thus, this transformation was used to treat the word frequency in both training sets.

3.2.3. Pattern Presentation in the Second Phase of Training

Although distributed representations established through the application of learning algorithms have several properties that are desirable for modelling language acquisition, there is also a less desirable property: new learning may cause the old knowledge to be completely forgotten when networks are trained sequentially. This effect is called Catastrophic Interference (McCloskey & Cohen, 1987).

Human learning is in most cases sequential, and this applies as well for the learning of a second language. The second language sometimes causes interference with the first language, but only seldom does the second language substitute for the mother tongue. At the same time, the most usual situation while learning a second language is parallel exposure to the first one. This is the kind of bilingual situation to be modelled with BAR 1. Thus, the training of the model should account for these two facts: sequential learning and simultaneous exposure to both languages.

Traditional cognitive models do not have special problems with sequential learning, because each concept is represented as an individual node. As already mentioned, distributed representations on the other hand do not deal very well with sequential learning, because each connection weight is involved in responding to many different input units. Thus, the way the back-propagation algorithm adjusts the weights in order to encode the desired response to a new input pattern will necessarily alter previously learnt responses to other input patterns.

If two training sets have to be learnt sequentially, the connection weights established after the first training will be changed during the second training.

As a consequence, the network only learns the second training set and the memory for the first set is destroyed.

McCloskey and Cohen (1987) tried to reduce the severity of the interference by carrying out a variety of manipulations. Among other things, they varied the number of hidden units and the learning parameter, overtrained the first training set, and froze the weights after the learning of the first set, but none of these variations really eliminated the catastrophic interference. One of the possible approaches, suggested by Rumelhart (McCloskey and Cohen, 1987) is the addition of context units to the input units. Although they do not claim that this approach will solve the interference problem, it might become less severe. Therefore, three context units were added to the input units in the BAR model, as already mentioned in the description of the architecture.

Another suggestion is given by Murre (1992). According to Murre the catastrophic interference primarily depends on the method of pattern presentation. He proposes that interference can be reduced by applying a random rehearsal method, which implies that the old patterns are trained together with the new ones.

This method was adopted in order to train the network with the set of English words once it had learnt the set of Dutch words. In each training epoch some of the new English words were presented together with some of the old Dutch words. Each word in this new training set was presented in random order. In this way English words were learnt by the network while at the same time Dutch words were not forgotten.

3.2.4. Software Package

The simulation was done on a UNIX workstation with a SPARC 10 processor.

The neural network was simulated by the PDP package (McClelland & Rumelhart, 1988), using the bp program which implements the back-propagation procedure. There are many programs which implement the Back Propagation algorithm, but the PDP package offers the possibility of changing the source code.

Two main modifications were made in the original bp program. The source code had to be changed in order to read the pattern file, which included the frequency specification for each word. Another option which that was added to the package is the computation of two separate error scores. As in the original model of Seidenberg & McClelland (1989), one error score is computed for the orthographic output and the other one for the phonological input.

Other programs were needed in order to run the simulation. The file with the set of words selected from CELEX contained the word-name, the orthographic representation, the phonological representation, and the frequency of the word. First, a program transformed this file into the pattern file that the input units could read: the training pattern file. This file contained the coded frequency of the word, the pattern name, the coded orthographic and phonological input to the network, and the coded orthographic and phonological output expected. The files to test the network were derived from this one by removing the expected output. Testing the network involves providing only the input patterns and registering the output given by the net.

After the simulation, the activation of the output units had to be transformed again to a readable word representation, in order to examine the learning of each particular word in its orthographic and phonological forms. Another program was written to do this decoding.

3.3. Version 1: Results and Analyses

3.3.1. Learning Accuracy

The monolingual as well as the bilingual training were conducted for 3000 epochs, in order to attain a steady error score. Before starting the monolingual training, the weights of the network's connections were randomly set. The monolingual training was started several times to check if the different weights randomly assigned could influence the error scores, and no large differences were detected after one hundred epochs of training. When the monolingual training phase concluded, the weights were saved in order to provide the starting weights for the bilingual training.

Figure 2.5 shows a graphical representation of the error scores of BAR 1 after both the monolingual and the bilingual training.

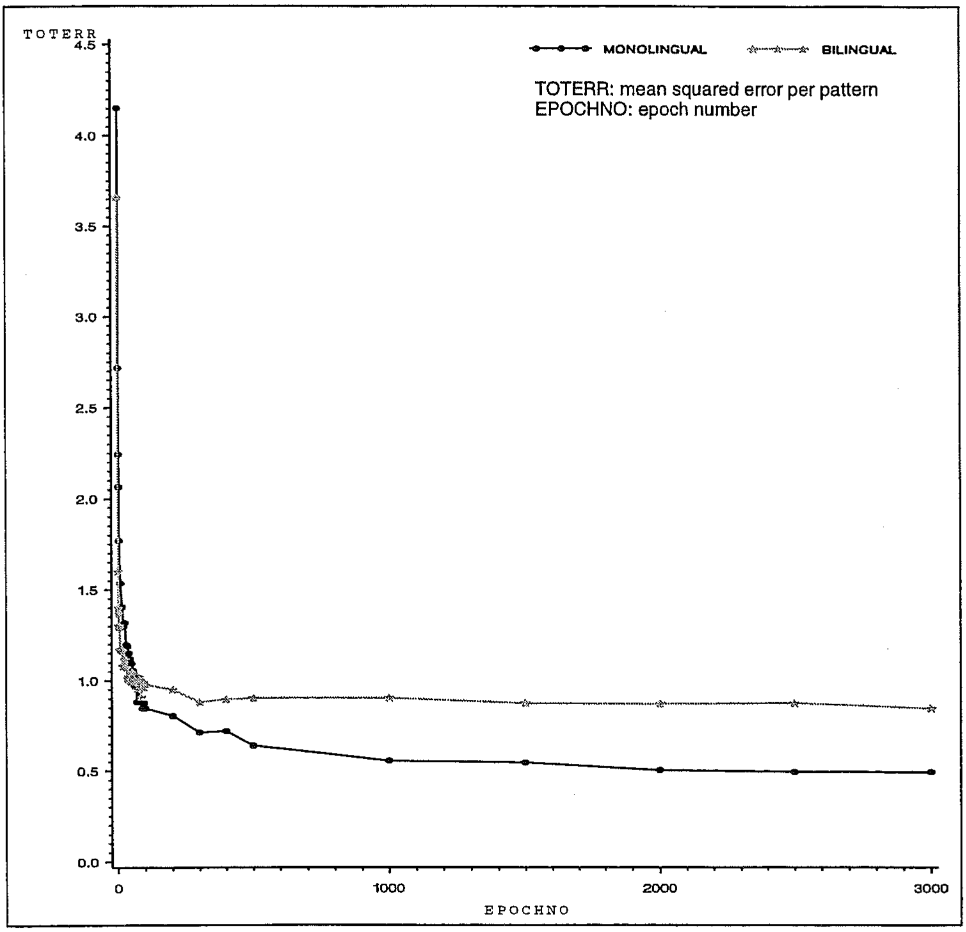

Figure 2. 5. BAR 1 Learning rates per epoch during Monolingual and Bilingual training

The total error score is the sum of squared errors between the actual output and the target output over all patterns. These error scores were determined for every one hundred epochs of training. In the figure, the error scores have been interpolated by straight lines in order to show their development. As can be seen in the figure, the network learns very fast during the first 100 epochs in both monolingual and bilingual training. After the epoch 1000 the error scores are rather steady: monolingual error scores converge to approximately 0.49

and bilingual error scores to approximately 0.85. The fact that in both training phases the error scores converge to a particular value confirms that word presentation based on word-frequencies does not result in large fluctuations of the error scores.

The next subsections are concerned with the different analyses applied to the results. In the first two subsections the error scores are further analysed for both monolingual and bilingual training. The aim of these analyses is to examine the effect of word frequency and word length on the learning of words, and to characterize the errors of the network's output in terms of orthographic and phonological accuracy. In the last two subsections, the internal representations built up by the network are examined in the light of experimental data.

3.3.2. Monolingual Training

In order to give a first impression about the quality of the performance of BAR 1 after monolingual training, the following table summarizes the percentage of the words learnt incorrectly. Three types of incorrect words are distinguished in the table, depending on to which representation of the word has been incorrectly learnt: Only Orthography, Only Phonology, and Orthography+Phonology. In order to find out the total percentage of orthographic errors and phonological errors, the category Orthography+Phonology is added, respectively, to the Only Orthography and

Only Phonology percentages. Thus it is possible to find out whether the network has more difficulties learning one type of representation or the other.

Dutch	Incorrect words (%)
Only Orthography	2.94
Only Phonology	3.37
Orthography+Phonology	1.15
Total Orthography	4.09
Total Phonology	4.52
Total	7.46

Table 2.1. Percentage of incorrect Dutch words after 3000 epochs of monolingual training

The percentages in Table 2.1 indicate that the performance for both types of representation is very similar (4.09% vs. 4.52%). The total number of words learnt incorrectly is 7.46%. In order to identify the factors causing these errors, several analyses were performed. These analyses are presented next.

Performance according to word frequency

As explained in section 2.1., during each epoch different words were presented to the network according to their word frequencies: words with high frequencies had a higher probability of being presented than words with low frequencies. Consequently, high frequency words should have been learnt before low frequency words, and should have been better represented in the initial stages of training.

In order to investigate the effect of frequency on learning accuracy, two subsets of words were extracted from the monolingual training set. The first subset contained 200 Dutch words with the lowest possible frequencies (6) and the second subset contained 200 Dutch words with the highest possible frequencies (from 122 to 1370). Table 2.2 shows the differences between the learning of low-frequency and high frequency words:

Dutch	Incorrect low-freq. words (%)	Incorrect high-freq. words (%)
Only Orthography	6.00	0.00
Only Phonology	3.50	0.00
Orthography+Phonology	2.00	0.00
Total Orthography	8.00	0.00
Total Phonology	5.50	0.00
Total	11.50	0.00

Table 2.2. Percentage of incorrect Dutch low-frequency and high-frequency words after 3000 epochs of monolingual training

The total percentage of incorrect words in the table shows that no errors were made in the representation of the 200 high frequency words, while 11.50% of the 200 low-frequency words were represented wrongly. This percentage is probably too high with respect to the expected performance of an adult Dutch speaker. It may indicate that the set of correspondences between Dutch orthography and phonology has not been properly learnt by the network.

The percentage of orthographically incorrect words is higher for low-frequency words than for the phonologically incorrect words (8.00% vs. 5.50%) and is twice the percentage of orthographic incorrect words for the total amount of words learnt (8.00% vs. 4.09%, see Table 2.1.). This fact can either indicate that the network has more difficulties learning the orthographic representation, or that the Dutch low-frequency words have a irregular orthography with respect the high-frequency words.

Performance according to word length

As described in section 3.1.3. (Coding), the coding used for the representation of the input information consists of two templates that have a syllabic structure (left-justified template: CCCVVCCCVVCCCVVCCC; right-justified template: VVCCC). These templates emphasize the beginning and the end of the words, because the left justified template codes a word from left to right; and the right justified template codes it from right to left, thus starting by the end of the word. These two features have some consequences for the encoding of the words that are summarized next.

The first consequence is that short words, as for example *een,* are represented twice; whereas long words have a chance of not being fully represented if they have more than 18 characters (i.e., they do not fit in the left-justified template), as for example *verantwoordelijkheid.*

Secondly, the syllabic structure described in these templates does not match exactly the syllabic structure of some Dutch words. In the Dutch language it is common to find groups of four or more consonants (for example, e<u>chtsch</u>*eiding*). The templates only admit combinations of three consonants and two vowels. Accordingly, a group of more than three consonants is coded as belonging to two syllables, skipping the vowels. This feature also causes long words to be incompletely represented. In the example given above, *echtscheiding* is a word of thirteen letters and it could be fully represented within the left template. But the syllabic structure forces to represent it as:

cccEvCHTvvSCHEIDcc

in the left justified template, and

IvNGc

in the right justified template. Only with the combination of the two templates can the word be fully represented. For other words the complete representation is not possible. Notice as well that three spaces are lost instantly when the word starts with a vowel.

The right justified template is not very useful for the representation of the words for BAR 1, since only two syllabic groups (VVCCC) are represented. As mentioned in Subsection 2.1.3., the choice of this template and its structure seemed to be related with the fact that McWhynney and Leinbach (1991) worked with the learning of past tense in English words, which in its regular form consists of the morpheme *-ed* added at the end of the infinitive form. The Dutch words that BAR 1 learned do not have such a regular structure and some words finish with a vowel group. Thus for Dutch words such as *knie*, the right-justified template codes only the last two vowels:

IEccc

The repetition of the last two vowels in the representation of the word is not very useful information for BAR I. Depending on the structure of each word, in some cases only vowels will be represented in the right justified template, in other cases the full last syllable will be represented, and in the case of short words, as already mentioned, the word will be entirely represented both in the left and the right-justified templates. As BAR 1 is supposed to learn spelling-to-sound correspondences, the irregularities in the representation of words are not useful information.

All these particularities of the coding can cause the network some difficulties in learning Dutch words, especially the longest ones. In order to examine this issue, an analysis similar to the one performed in last subsection was conducted. A subset with the 200 shortest words and a subset with the 200 longest words were extracted from the training set and the percentage of errors for each subset were compared. Results are displayed in the following table:

Dutch	Dutch shortest words (%)	Dutch longest words (%)
Only Orthography	1.00	15.00
Only Phonology	0.00	14.00
Orthography +Phonology	0.00	8.00
Total Orthography	1.00	23.00
Total Phonology	0.00	22.00
Total	1.00	37.00

Table 2.3. Percentage of incorrect Dutch short and long words after 3000 epochs of monolingual training

Table 2.3 shows that the performance on short words is almost perfect, but long words, as predicted, are not learnt properly. From the set of 200 long words, 37% of the words are not correct. As there are not big differences in the type of errors produced by the network, it can be concluded that the problem is the word length, and this can be attributed to the incomplete coding for long words. These results confirm that the two-templates scheme is not the most appropriate for the actual purposes of this project.

3.3.3. Bilingual training

As was done in the case of monolingual training, the first results examined after the bilingual training were the total percentages of incorrect words. These percentages are summarized in table 2.4 for both English and Dutch languages.

Dutch & English	Dutch Incorrect words (%)	English Incorrect words (%)
Only Orthography	7.60	6.19
Only Phonology	7.60	5.55
Orthography+Phonology	7.92	6.24
Total Orthography	15.52	12.43
Total Phonology	15.52	11.19
Total	23.11	17.98

Table 2.4. Total percentages of incorrect words in Dutch and English after 3000 epochs of bilingual training

The percentages of incorrect words obtained after 3000 epochs of bilingual training are not very different from one language to the other, and they are comparable also across modalities (orthographic and phonological). The performance on Dutch is poorer after the bilingual training (23.11% of incorrect words) than it was after the monolingual training (7.46%), indicating the effects of interference caused by learning the new set of words. The error score obtained after the second phase of training (0.85, see fig. 5) is caused by errors in both languages, although the performance in English is slightly better. Next, the same analyses performed after the monolingual training are presented, in order to highlight the factors that might cause difficulties in learning the two sets of words: word frequency and word length.

Performance according to word frequency

One of the main problems that could be encountered in the phase of bilingual training was, as already mentioned, the interference of the new patterns in the learnt patterns. After one epoch of bilingual training the sets of high and low frequency Dutch words were tested again, in order to check the effect of the interference produced by the learning the second set of features corresponding to English. Table 2.5 shows the percentage of incorrect words after one epoch of bilingual training.

Dutch	Incorrect low-freq. words (%)	Incorrect high-freq. words (%)
Only Orthography	11.00	6.00
Only Phonology	21.00	13.00
Orthography+Phonology	21.00	2.00
Total Orthography	32.00	8.00
Total Phonology	42.00	15.00
Total	53.00	21.00

Table 2.5. Percentage of incorrect Dutch low-frequency and high-frequency words after 1 epoch of bilingual training

The percentage of errors in Dutch words after one epoch of bilingual training has increased considerably. When compared to the performance showed in Table 2.2, the number of incorrect low-frequency words has increased from 11.5% to 53%; and the number of incorrect high-frequency words has increased from 0% to 21%. It is noticeable that the network makes more mistakes in representing the phonology than in representing the orthography of the words, unlike the results displayed in the former tables. This effect could be due either to the particular coding used for phonological representation, or to the fact that the main differences between these two languages lie in the phonological representation.

At the end of the bilingual training the effects of interference were examined again. The percentages of incorrect English low frequency and high frequency words were added to the results. As in the case of Dutch words, two subsets of 200 English words each, for high and low frequencies, were selected from the bilingual training set. The results are shown in Table 2.6.

Dutch & English	Incorrect Dutch low-freq. words (%)	Incorrect Dutch high-freq. words (%)	Incorrect English low-freq. words (%)	Incorrect English high-freq. words (%)
Only Orthography	10.00	7.00	6.00	2.50
Only Phonology	13.00	3.50	19.00	5.50
Orthography+Phonology	12.00	2.00	8.00	1.00
Total Orthography	22.00	9.00	14.00	3.50
Total Phonology	25.00	5.50	27.00	6.50
Total	35.00	1.50	33.00	9.00

Table 2.6. Percentage of incorrect low-frequency and high frequency words after 3000 epochs of bilingual training

Comparing the results of this table to those of the previous one, it turns out that the percentage of incorrect Dutch low-frequency words has decreased from 53% to 35%, and the percentage of incorrect Dutch high-frequency words has decreased from 21% to 1.50%. Although the network recovered from many of the errors, the number of errors in Dutch words after 3000 epochs of bilingual training still remains higher that the number of errors after the monolingual training.

Comparing the percentages across modalities for both languages, it is remarkable that the network makes fewer errors in the orthographic representation of English words (14% for low-frequency words and 3.50% for high frequency words) than for Dutch words (22% for low-frequency words and 9.00% for high frequency words). Within the English words, the percentage of orthographic errors is also lower than the percentage of phonological errors for both high- and low-frequency words. In general, the percentages are lower for the sets of English words than for the sets of Dutch words. It is very likely that these differences are due to the interference of English words in learning Dutch words.

Performance according to word length

As in the case of monolingual training, a set of long and short words was tested for both Dutch and English words after bilingual training. The results are shown in Table 2.7.

Dutch & English	Dutch shortest words (%)	Dutch longest words (%)	English shortest words (%)	English longest words (%)
Orthographic	1.00	17.00	0.00	9.00
Phonology	0.00	20.00	1.00	33.00
Orthographic+ Phonological	1.00	43.00	0.00	30.00
Total Orthography	2.00	60.00	0.00	42.00
Total Phonology	1.00	63.00	1.00	63.00
Total	2.00	70.00	1.00	72.00

Table 2.7. Percentage of incorrect Dutch and English short and long words after 3000 epochs of monolingual training

There are no big differences between Dutch and English with respect to the percentages of incorrect words, either long or short. The percentage of errors obtained for long words is very large. In the case of Dutch words, the total percentage of incorrect long words (70%) is practically double that obtained after monolingual training (37%). It is remarkable that the largest number of errors is made on long words for both languages. This fact seems to indicate again that the coding used is not suitable for learning the long words.

3.3.4. Analysis of the Internal Representations

As mentioned in the introduction of this section, the aim of the BAR model is to offer a new model of bilingual lexical organization that is able to learn new words. At the same time, the model is intended to account for the cognate effect. Our claim is that with a distributed representations model this effect is better explained than with a local representation model. The distributed representations of BAR 1 should capture the orthography and phonology interactions in such a way that similar words are represented with a similar pattern of activation through the hidden units. In order to see if BAR 1 actually represents these features in its hidden layer, we used a clustering technique to compare the different patterns of activation of the hidden units for some words. Clustering techniques are used to place objects into groups or clusters suggested by the data, not defined a priori, such that objects in a given cluster tend to be similar to each other in some sense, and objects in different clusters tend to be dissimilar. The Hierarchical Cluster Method was used to analyse the internal representations in BAR 1 after bilingual training .

The set of words used for this analysis was selected from the list shown in the appendix of De Groot & Nas (1991), which were also present in the training set for the network (see Table 2.8). The set includes similar and dissimilar words both within and between languages, hence it is a good set to examine the patterns of internal representation.

prime	TARGET	prime	TARGET
spiegel	MIRROR	lichaam	BODY
gezicht	FACE	winkel	SHOP
hart	HEART	varken	PIG
haar	HAIR	geld	MONEY
muziek	MUSIC	grap	JOKE
hel	HELL	wortel	CARROT
broer	BROTHER	vleugel	WING
appel	APPLE	kantoor	OFFICE
vorm	FORM	broek	TROUSER
midden	MIDDLE	mes	KNIFE
pond	POUND	konijn	RABBIT
dief	THIEF	horloge	WATCH
bal	BALL	paard	HORSE
klok	CLOCK		

Table 2.8. Set of words used for the Hierarchical cluster Analysis (BAR 1)

Each word in this set is represented by a different pattern of activation of the hidden units of BAR 1. The clustering analysis compares these representations to each other, and organises them in a two dimensional space according to their similarity. The distance at which the words are clustered indicates their similarity/dissimilarity: when two words are similar, the distance between them is very short. The results of single linkage clustering are shown in figure 2.6.

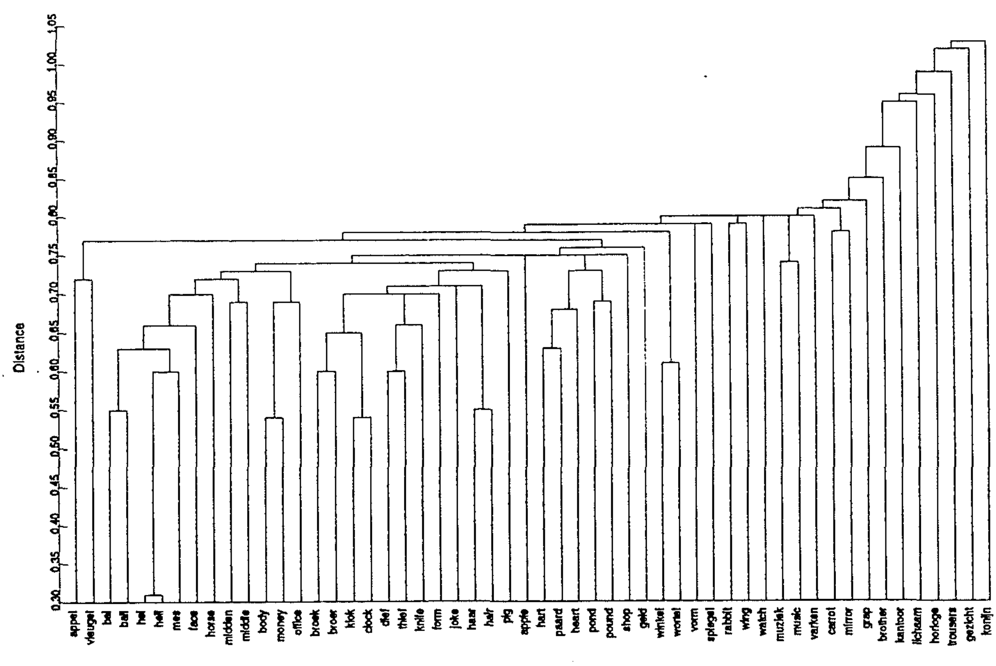

Figure 2.6. Single-Linkage Clustering results

Obviously, the results of this clustering analysis are only approximate, because the resulting clusters would possibly be very different if the whole set of trained words was clustered.

As expected, words are classified into coherent groups when they have similar forms and words are classified into different groups when they have different forms. The cluster distances range from 0.31 to 1.03. Below the mean distance (0.67) the following words are clustered:

(((hel hell) mes)(bal ball))
((klok clock)(broek broer))

(body money)
(haar hair)
(dief thief)
(winkel wortel)
(hart paard)

The brackets indicate the order in which the clusters are created. For example, first the clusters (*hel hell*) and (*bal ball*) are created. Subsequently a new cluster is formed containing the word *mes* and the (*hel hell*) cluster, and finally the cluster (*bal ball*) and the former cluster are clustered together.

The words *konijn*, *gezicht*, *trousers*, *horloge*, *lichaam*, *kantoor*, and *brother* are the most dissimilar words within this set, and this dissimilarity is reflected in the large inter-cluster distances.

The results of this analysis indicate that the patterns of activation of the hidden units are similar for similar words, and very different for dissimilar words. The internal representations built by BAR 1 correspond the expectations, and enable the model to account for the cognate effect, as mentioned in the introduction of this subsection.

3.3.5. BAR 1 and Experimental Data

The final analysis on BAR performance was done by comparing the results of the simulation with experimental data using the same set of words. For this comparison we used the results obtained by De Groot & Nas (1991, exp. 4)[1]. Although the error scores from the learning rates of BAR 1 were too high to consider the model to be fully developed, it was interesting to check if it concurred with some of the trends of human data.

The Spearman rank correlation method was used to compare two sets of completely independent data: the clustering results from BAR 1 and the reaction times obtained in the experiment of De Groot and Nas (1991, exp. 4). The data are not directly comparable because the results of the simulation are expressed in minimal cluster distances, whereas the results of De Groot & Nas experiment are expressed in milliseconds (see Table 2.9. with both measures).

1. The author wishes to thank Annette De Groot for kindly providing the raw data of her experiment.

Prime	Target	Mean reaction time (ms)	Minimal cluster distance
spiegel	MIRROR	559.158	0.8060
lichaam	BODY	510.789	0.9548
winkel	SHOP	517.158	0.7795
varken	PIG	545.556	0.8047
geld	MONEY	498.211	0.7598
grap	JOKE	554.579	0.8171
wortel	CARROT	632.941	0.8060
vleugel	WING	528.278	0.7963
kantoor	OFFICE	526.333	0.8936
broek	TROUSER	642.778	0.9928
mes	KNIFE	566.437	0.7355
konijn	RABBIT	558.368	1.0299
horloge	WATCH	496.211	0.9556
paard	HORSE	491.000	0.7507
gezicht	FACE	510.789	1.0205
hart	HEART	486.053	0.6807
haar	HAIR	471.611	0.5626
muziek	MUSIC	514.778	0.7361
hel	HELL	516.333	0.3114
broer	BROTHER	500.053	0.8488
appel	APPLE	467.158	0.7657
vorm	FORM	554.125	0.7866
midden	MIDDLE	530.167	0.6934
pond	POUND	525.211	0.6858
dief	THIEF	537.778	0.6012
bal	BALL	524.500	0.5507
klok	CLOCK	495.842	0.5400

Table 2.9. Reaction Times (De Groot And Nas, 1991, exp. 4) and Cluster distances (BAR 1)

The Spearman rank correlation coefficient that is associated with the mean reaction times and the minimal cluster distances has a small positive value (rs = 0.32). This implies that there is a small tendency of the two ranks to be similar, which is not statistically reliable ($p > 0.050$).

3.4. Discussion

The results obtained after the first simulation of the Bilingual Access Representations model (BAR 1) were analysed in different ways. The first analysis was of the error scores of the network after both the monolingual and the bilingual phases of training. Secondly, each phase of training was analysed independently. In both phases, the factors word frequency and word length were examined. For the bilingual training the internal representations built up by the network were also studied with a clustering technique, and finally we compared the performance of BAR 1 with experimental data.

3.4.1. Error Scores

The error scores obtained after both monolingual and bilingual training indicate that after the first simulation the results are satisfying. Nevertheless, the error score obtained (0.49) after monolingual training suggests that the correspondences between orthography and phonology for Dutch language have not been learnt completely. Regarding the results after the bilingual training, the bilingual error score is higher (0.85). As this error score stands for both languages, the rest of the analyses done on these results helped to inquire whether the high error score was due to one of the languages or both.

3.4.2. Monolingual Training

The errors observed after the monolingual training can be summarized as follows: first, better performance on high-frequency words than on low-frequency words; second, the almost perfect performance on short words and the very poor performance on the long words. From these observations, it can be concluded that the high rate of errors is mostly due to the bad learning of the long words. The possible reasons for these errors of the network are the following:

1. Small size of the training set. The Dutch training set was composed of 3738 words, approximately the same size as Seidenberg & McClelland's (1989) training set. Seidenberg & McClelland (1990) remark that this size is maybe too small to provide a perfect learning of the set of orthography/phonology correspondences, and that can apply for BAR 1 as well. The fact that low-frequency words are not learnt by the network might indicate that some particular features of Dutch orthography/phonology correspondences were not presented enough times during the training phase. Consequently, the next simulation of BAR should be trained with a bigger set of Dutch words.

2. The coding used. The coding used might be responsible for two different kinds of errors:

2.1. Spelling-to-sound correspondences: The examination of the spelling errors produced seem to indicate that the phonological coding generates internal representations very unlike from human errors, as for example *voorterp* instead of *voor<u>w</u>erp*. A way to approximate network's performance to human performance is to give the network phonological information based on phonological features such as sonority and articulation points. With this information it is expected that the network would confuse phonemes that are similar rather than learning some random changes.

2.2. Biased errors due to the double coding of the end of the words. As was mentioned earlier in this chapter, the coding used for BAR 1 was adapted from McWhynney and Leinbach (1991). It was originally designed for a model on the learning of past tense. As the morphological information for the verb tense is located at the end of the word, McWhynney and Leinbach emphasized this part of the word by encoding it twice. As the purposes of the BAR model are different, the encoding in two templates is not necessary. BAR has to learn the proper representation for a whole word, and the coding used implies that very long words are not fully represented and the network does not learn these words properly. This feature is not desirable at all.

Summarizing, the poor performance after the monolingual training seems to be due partly to the small size of the training set, and partly to the coding used. On the one hand, the lack of phonological information results in poor learning (i.e., random substitutions of phonemes), and on the other, the template encoding seems to be responsible for the high error scores (i.e., long words are not properly learnt).

3.4.3 Bilingual Training

After the bilingual training, poorer performance is obtained. The error scores are higher, and the different tests applied to the results (performance in high and low frequency words, and performance on short and long words) show that this error score is due to bad performance in both languages. Basically, the same problems indicated above for the monolingual training apply to bilingual training as well. The fact that the performance of the network in Dutch is worse after the bilingual training than after the monolingual training indicates that the effects of interference have not been avoided completely.

In spite of the above mentioned problems, BAR 1 shows good performance in terms of internal representations: from the cluster analysis it is obvious that similar words both within and between languages have a similar internal representation, and there is a trend in the Spearman coefficient indicating the correlation of this representation with response times in human experiments.

A way to improve the performance of the network with respect to learning Dutch is to enlarge the size of the Dutch training set and reduce the size of the English training set. For the bilingual training this implies that the exposure of the network to both languages better models the situation of an adult Dutch speaker learning English as a second language. Another aspect to be revised is the phonological coding used by the network.

The conclusion following form the BAR 1's simulation is that distributed representations are able to cope in a single lexical level with two languages, but that the performance of the network has to be improved in several respects. In the next section BAR 2 is presented, a new version of the model which incorporates the suggested improvements.

4. Bilingual Access Representations Model: Version 2 (BAR 2)

The analyses of the performance of BAR 1 indicated that the model met the expectations concerning the internal representations for lexical entries of the two different languages. Nevertheless, in terms of learning, this performance was considered too poor and two improvements were suggested, both related to the coding used for the input and output of the information.

BAR 2 was designed to meet the following two requirements. On the one hand, the template scheme that contained the codes for the words, both orthographic and phonological, was changed. On the other, the phonological coding reproduced some articulatory features, in an attempt to avoid random substitutions of phonemes by the network.

BAR 2 also differs from BAR 1 in terms of training. As the error score in BAR 1 did not show major changes after the training epoch 1000, both monolingual and bilingual training phases were conducted for 1000 epochs. Another difference is the size of the training sets. While for BAR 1 both Dutch and English training sets had the same size (around 4000 words), the Dutch training set used for BAR 2 contained about 8000 words and the English training set about 2000 words. For the monolingual training, a bigger set of words allowed better learning of low-frequency words. For the bilingual training, the difference in size between the two sets implied less interference of the learning of the new words. Moreover, the proportion of words in each language was considered to be closer to that of a Dutch adult speaker who has learnt English as a second language. This was an important aspect, since the way (see Sebastian-Galles, 1991, for Spanish). On the contrary, Chinese readers seem to use articulatory codes as well as analogical codes to decode their logographs (Seidenberg, 1985). Overall the evidence suggests that there is an universal pattern for reading using both ways of accessing the lexicon (visual and articulatory), independently of orthography depth (Besner, 1987). Thus, it can be assumed that different orthographies should not affect the way words are represented in bilingual lexicon.

Recently, speech segmentation has attracted much research interest. As speech is a continuous process, it is assumed that the generation of an internal representation of auditory input is serial. Speech segmentation is then needed in order to organize this input into meaningful units. Although for years this process was supposed to be universal and so all hearer/speakers would do it in the same way, Cutler, Mehler, Norris & Segui (1986; 1989; 1992) showed that different languages settle different boundaries for their phonological processing. Their main finding is that French listeners use syllabic segmentation, while English listeners rely on other speech features such as stress, using strong syllables as a cue. In their experiments with bilinguals (Cutler *et al.*, 1992) the authors found that bilingual hearers/speakers rely on their first language for further segmentation. Even balanced bilinguals very proficient in their two languages use one of the two systems for speech segmentation. Similar research has been carried out for Spanish and Catalan listeners (Sebastian,N.; Dupoux, E.; Segui, J. & Mehler, 1992), for Japanese (Otake, Hatano, Cutler & Mehler, 1993), and Dutch (Van Zon & De Gelder, 1993), showing different types of segmentation according to the language. This implies that bilingual phonological processing of speech is actually monolingual. Speech segmentation is a language-specific factor and has to be taken into account when modelling lexical access in bilinguals.