Learning to predict the phonological structure of English loanwords in Japanese

Alan D. Blair
School of Computer Science & Engineering
University of New South Wales
2052, Australia
blair@cse.unsw.edu.au

John Ingram
Department of English
University of Queensland
4072, Australia
jingram@cltr.uq.edu.au

Abstract

Loanword formation seems to provide a good test bed for the growing field of computational phonology, since it occurs in a more tightly controlled environment than other language processing tasks. We show how feedforward neural networks and decision trees can be trained to predict the phonological structure of English loanwords in Japanese, and compare the performance of the two paradigms. In each case the system produces a phonemic representation of the Japanese form, after receiving as input the phonological feature matrix of the current and surrounding phonemes. The performance is improved with the inclusion of information about the stress pattern, orthography of reduced vowels and location of word boundaries.

1 Introduction

When words are borrowed from one language into another, they typically undergo systematic changes to make them conform to the phonology of the borrowing language. These changes are rather complex and involve many interacting factors. Phonologists have traditionally sought to construct symbolic algorithms for these kinds of tasks, either with or without reference to a framework of learnability theory. This symbolic approach came under strong challenge from empirical connectionist models of language processing in the late 1980's and early 1990's (Gupta & Touretzky, 1994). The debate between these competing paradigms, or the search for some suitable hybrid, continues.

In an earlier study (Blair & Ingram, 1998) we showed how a feedforward neural network can be trained to predict the phonology of loanwords borrowed from English into Japanese. The issue of loanword formation and the architecture of this network are reviewed in sections 2 and 3, respectively. In section 4 an analysis of these results is presented, which brings to light some shortcomings of our original network. In section 5 we explore a number of possible improvements arising from this analysis, and examine how the network performance is affected by each of these modifications. The performance of the neural network is also compared with that of a benchmark decision tree algorithm called C4.5 (Quinlan, 1992). The paper concludes with a discussion of the results and an outline of plans for future work.

2 Loanword formation

Loanword formation provides a good environment for studying phonological structure and mechanisms, because it is more restrictive, and hence better controlled, than other language processing tasks (which involve a host of lexical, morphological, syntactic or pragmatic influences). Silverman (1992) has described loanword formation as fundamentally a two-stage process: the first stage yields a parsing of the phonetic input into segmentally organised phonetic feature bundles, interpretable as segmental targets in the borrowing language; in the second stage, these segmental targets are parsed into phonological structures (syllables, mora, feet, etc.) compatible with the word-prosody of the borrowing language.

Loanwords from English into Japanese are particularly interesting from both a theoretical and a practical point of view. On the theoretical side, while the segmental mapping from English to Japanese is relatively straightforward, their respective word-level prosodies are strikingly different, providing ample opportunity to observe prosodic re-structuring in loanword formation. On the practical side, the Japanese language has borrowed thousands of words from English, providing an abundant source of data, and there is a perennial need for converting the names of people, places, companies, products and creative works from English into Japanese in a consistent and natural manner.

While the segments and phonetic features of English words tend to be remarkably well preserved by the process of loanword formation, the resulting Japanese word forms are so completely transformed in their prosodic structure that English listeners almost invariably fail to recognise their English sources, when loanwords are presented to them as isolated words carefully spoken by a native speaker of Japanese (Ingram, 1998). The main factor underlying poor recognition of the English source words appears to lie in the extensive resyllabification, involving vowel epenthesis, which is required to parse the segmental input into Japanese prosodic frames. Some examples are given below:

orthographic	phonemic	Japanese
Olympic	olimpik	oriɴpikku
truck	trak	torakku
cut	kat	katto
cud	kad	kado
cart	kart	kaato
cat	kæt	kyatto

Many English phonemes do not occur in Japanese, and many combinations of phonemes occurring in English are forbidden in Japanese (even though the individual phonemes may occur separately). Japanese syllables almost invariably consist of a vowel by itself or a single consonant followed by a vowel, thus lacking complex onsets and codas. To maintain a faithful representation of the segmental structure of the English source word, extensive use is made of epenthetic vowels — chiefly /u/,/o/ or /i/, depending on the context. The temporal structure of the source word is converted to Japanese moraic timing. English voiceless obstruents sometimes become geminate (two mora) stops in Japanese — for example, /t/ becomes /tto/ at the end of /kyatto/. Such gemination is most likely to occur after a short, stressed vowel but there are many exceptions to this rule. In general, it is a difficult problem to predict when gemination will occur and when it will not.

The Japanese language has only five distinct vowels (/a/,/i/,/u/,/e/,/o/) while English has at least eight (excluding diphthongs). On the other hand, vowel length is a more salient feature in Japanese than it is in English. In the process of loanword formation, each vowel in the original word will tend to be replaced by a similar vowel among those available in the borrowing language. However, the length and quality of this vowel is not simply determined by the original vowel, but is also dependent on the surrounding phonemes and a combination of other factors, including the temporal structure and stress pattern of the source word.

The processes involved in loanword formation depend on a number of factors interacting in complex ways. It is generally true that the sound structure of loanwords may be seen as the result of filtering the phonetic form of a foreign word through the phonological system of the borrowing language. However, it is also the case that loanwords typically posses characteristics that mark them off from the native vocabulary, by their exceptional behaviour with respect certain otherwise general restrictions or rules of sound structure (e.g. in Japanese, word forms from the Native (Yamato) or Sino-Japanese strata of the lexicon do not permit single /p/ consonants to occur between vowels, but only as geminates:

nippoN	[Japan]	*nipoN
kappa	[river-imp]	*kapa

But this restriction does not apply to the large portion of the vocabulary of more recent foreign origin, including English loanwords (Ito & Mester, 1996):

peepaa [paper]

Moreover, loanwords (recent borrowings in particular) are subject to additional sources of variation in pronunciation and written form, depending on the linguistic sophistication of users and the context of usage. Competing phonological constraints in the borrowing language may result in native speaker indecision as to how to construe or parse the phonetic input of a foreign word form. While the process of loanword formation is systematic, it is also subject to a degree of indeterminacy. There are various ways to estimate the extent of this indeterminacy – for example by asking native listeners to transcribe into their native orthography a list of foreign spoken words or word-like stimuli (Takagin & Mann, 1994). The form provided in a dictionary or database is not necessarily the only acceptable option – a fact which must be taken into account in the evaluation of any automated system for predicting loanword phonology.

3 Representation and timing issues

3.1 Representation

Before determining an architecture for our system, it is necessary to choose an appropriate representation for the inputs and outputs. For the outputs, we employ a *phonemic* representation, with each individual output directly corresponding to one of the 20 phonemes in the Japanese language. Although a similar phonemic representation could in principle be used for the (English) input of the system, we choose instead to employ a *featural* representation for the input, in which each phoneme is encoded as a combination of thirteen (binary) linguistic features: syllabic, continuant, sonorant, sibiline, nasal, voiced, labial, coronal, high, back, low, affective and lateral (see Appendix). This featural representation provides a number of advantages over its phonemic counterpart: (1) it reduces the total number of inputs – thereby simplifying the structure of the resulting architecture, (2) features often influence the form of loanwords in a systematic way, thus facilitating generalization, and (3) the same word is often rendered differently in different dialects of the source language (for example, British vs. American English) and the featural representation is less sensitive to these differences than a direct phonemic encoding.

An analogy can be drawn here with the NETtalk system of (Sejnowski & Rosenberg, 1987) which learned to predict the phonology of English words (expressed in terms of phonetic features) from their orthographic representation. We may ask whether a system could learn to perform the *opposite* task — that of predicting the orthography of a word from its phonological features. This question has not aroused much interest as it stands, for two reasons: first, many of the world's languages employ a strictly phonetic writing system, for which it is trivial to convert phonology to orthography; second, most of the other languages use writing systems for which the task is messy and ill-defined, due to a lack of systematicity and a very large number of homophones. Predicting the phonology of loanwords in a language like Japanese, however, is in some sense analogous to the task of transforming phonology to orthography, since the orthography of the loanwords (in katakana) can easily be derived from their phonemic transcription. In other words, the task we are studying can be loosely thought of as the *reverse* of the NETtalk task, adapted to a new setting in which it becomes both challenging and of practical importance.

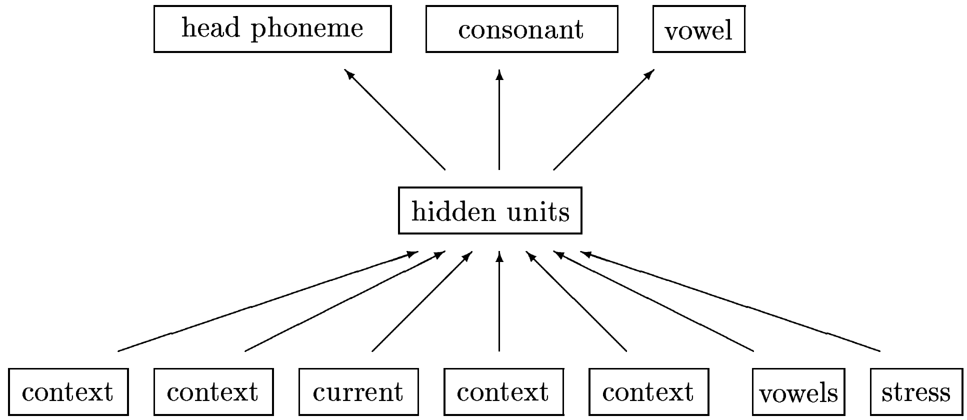

Figure 1: Neural network architecture.

3.2 Context and timing

The neural network used for our preliminary studies had 65 inputs, 20 hidden units and 53 outputs (see Figure 1). The 65 inputs are divided into 5 groups of 13 inputs, which are used to encode the phonological features of the current phoneme, the two preceding phonemes and the two following phonemes.

Input and output phonemes do not always correspond on a 1-to-1 basis. In some cases a phoneme may be deleted, or it may have a consonant and/or a vowel appended to it. In order to allow for these possibilities, we divided the outputs of our network into three groups. The first group has one output for each possible Japanese phoneme (consonant or vowel); the second group has one output for each possible consonant; the third group has one output for each possible vowel. Each group has one additional output representing the "empty" phoneme /_ /. Since there are 20 consonants and 5 vowels in Japanese, the total number of outputs is 26+21+6=53. For example, consider again the English word "cat" which has the phonemic representation /kæt/ and becomes /kyatto/ in Japanese. The network views this example as three separate training items:

$$\begin{array}{ccc} \_\,\_\,\mathbf{k}\,æ\,t & \rightarrow & k\,y\,\_ \\ \_\,k\,\mathbf{æ}\,t\,\_ & \rightarrow & a\,\_\,\_ \\ k\,æ\,\mathbf{t}\,\_\,\_ & \rightarrow & t\,t\,o \end{array}$$

This means that the network, when presented with the features encoding the input /__kæt/, should be trained to produce an activation of 1.0 for the /k/ output of the first group, the /y/ output of the second group and the /_/ (empty) output of the third group (and an activation of 0.0 for the other 50 outputs). When it comes to the testing phase, within each group the output with the largest activation is selected, and this determines the three-phoneme sequence chosen by the network to correspond with the current input phoneme.

3.3 Data

In order to minimise the effect of dialectal differences, we compiled a database of 1100 loanwords from a dictionary of neologisms (Bailey, 1962) which contains mostly loanwords that were borrowed into Japanese from American usage during a comparatively short period of time in the post-war era. To ensure consistency, the English phonemic transcriptions of these words were obtained from the Carnegie Mellon Pronouncing Dictionary (ftp://ftp.cs.cmu.edu/project/fgdata/dict/) which generally reflects American rather than British pronunciation.

4 Network training and preliminary analysis

4.1 Training

Networks were trained by back-propagation (Rumelhart et al., 1986) with a learning rate of 0.01 and a momentum of 0.9. The cross-entropy minimization criterion was used. Each of 11 networks was trained on 1000 words from the database, and tested on the other 100 words. Each word occurred in the test set of exactly one network. The 1100 words in the database had an average of 8.8 phonemes per word, making a total of 9660 input phonemes. Each of these input phonemes can produce output consisting of a head phoneme (group 1 outputs) plus an optional added consonant (group 2) and/or added vowel (group 3).

After 30 epochs of training, the networks achieved a combined error rate on the training and test set, respectively, of 7.4% (resp. 9.9%) for the head phoneme, 1.1% (resp. 2.0%) for the added consonant, and 1.1% (resp. 1.6%) for the added vowel, making a total error rate of 9.6% (resp. 13.5%). Note that in our data set, the head phoneme was nonempty 97% of the time, while the added consonant and added vowel were nonempty only 4% and 17% of the time, respectively, so we should not be surprised that the network error is much smaller for the latter two groups.

4.2 Analysis

An analysis was made of the errors produced by this initial network. The four most common error categories are summarised in Table 1; a more thorough linguistic analysis may be found in (Blair & Ingram, 1998).

Table 1. Summary of Error Categories

schwa vowel colouring	27%
vowel length	19%
obstruent gemination	11%
vowel epenthesis	7%
other	36%

By examining the errors in these common categories we deduced that the phonological features alone were not providing the network with all the information necessary for predicting the correct output. We identified three additional sources of information that might be useful to the network:

1. orthography of reduced vowels
2. location of word boundaries
3. stress pattern

The following section describes our attempts to reduce the number of errors by providing this kind of information to the system through additional inputs. The performance of the neural networks is also compared with that of decision tree learners.

5 Modifications and comparison of architectures

Based on the analysis of the previous section, a number of neural networks were trained using additional inputs, as outlined in the following subsections. These networks were generally trained for 30 epochs, although a few were trained for 50 or 100 epochs. For consistency, we always compare different networks by reporting the error rate, on the test set, after 30 epochs. The results of these experiments are outlined in Table 2. (Unfortunately, due to the large amount of data and long training times, it was not possible to perform each experiment multiple times with different initial weights.)

5.1 Comparing symbolic and connectionist models

Learning systems for linguistic tasks can be broadly divided into two main categories: *connectionist* models, such as the neural networks described in section 4, and *symbolic* models, such as decision trees or rule-based systems. Broadly speaking, connectionist systems tend to perform well on tasks requiring a classification which combines a large number of factors in an approximately linear fashion, while symbolic models tend to be more appropriate for tasks which rely on a smaller number of variables but combine them in a more nonlinear way.

In order to make a comparison between symbolic and connectionist models for the task of predicting loanword phonology, a benchmark decision tree algorithm called C4.5 (Quinlan, 1992) was used as a point of comparison for the neural network models. In each case, three separate decision trees were generated to predict the head phoneme, added consonant and added vowel, using the same inputs as the corresponding neural network.

As can be seen from the top row of Table 2, C4.5 achieves a combined error rate of only 11.5% compared with 13.5% for our original neural network. Moreover, with the inclusion of additional inputs (described in the following subsections) C4.5 continues to outperform the neural networks by 2 to 3 percentage points, suggesting that a symbolic approach may be better suited to this task than the connectionist alternative.

Table 2. Summary of Results

inputs	head		consonant		vowel		total	
	net	tree	net	tree	net	tree	net	tree
F	9.9	8.1	2.0	1.8	1.6	1.5	13.5	11.5
FS	11.5	8.1	1.7	1.5	1.6	1.5	14.8	11.1
FB	10.3	7.2	1.4	1.4	1.2	1.1	12.9	9.6
FBS	10.1	7.3	1.3	1.4	1.1	1.0	12.5	9.7
FO_1	8.8	5.7	1.7	1.8	1.7	1.5	12.2	9.1
FO_1S	8.8	5.7	1.9	1.5	1.8	1.5	12.5	8.7
FO_1B	8.3	5.2	1.4	1.4	1.3	1.1	11.0	7.6
FO_1BS	8.5	5.2	1.5	1.4	1.3	1.0	11.3	7.6
FO_2	8.8	6.0	1.6	1.6	1.7	1.5	12.1	9.1
FO_2S	8.0	6.1	1.7	1.5	1.7	1.4	11.5	9.0
FO_2B	7.8	5.5	1.4	1.2	1.1	1.0	10.3	7.6
FO_2BS	7.5	5.5	1.5	1.3	1.1	1.0	10.1	7.6

F : feature matrix of current and surrounding phonemes
O_1 : extra inputs for vowel orthography
O_2 : schwa replaced by orthographic vowel
B : word boundaries indicated by /_/
S : stress of current or immediately preceding vowel

5.2 Reduced vowels and orthography

The biggest problem for our original network (accounting for 27% of all errors) had to do with the colouring of reduced vowels. English has a reduced vowel (called the *schwa*) for which there is no Japanese equivalent. It seems that, when confronted with a schwa, Japanese speakers often look to the orthography of the English word to help them choose an appropriate vowel with which to replace it. This makes the task extremely difficult for the system to learn, since it has no direct access to the English orthography, but can only try to guess at it — based on the statistics of the input data.

We devised two different methods for presenting the orthography of the

English vowels to the system. The first method (labeled O_1) involved adding five additional inputs, to represent the five vowels {a,e,i,o,u}. Whenever the current phoneme is a vowel, one of these inputs is set to 1.0. With the second method (O_2), whenever a schwa was encountered in the input, its feature matrix was replaced with that of the corresponding orthographic vowel. Additionally, a new feature was added to the input feature matrix, to indicate whether or not the phoneme is a reduced vowel. As can be seen from Table 2, both the O_1 and O_2 representations of vowel orthography provide an improvement in performance of at least 1.2 percentage points for the neural network, and at least 2.0 percentage points for the decision tree learner. In the case of the neural network, although the O_1 and O_2 representations achieve about the same level of performance when used in isolation, the O_2 representation provides an additional improvement of about one percentage point over its O_1 counterpart, when used in combination with either of the other modifications described below (B & S). The reasons for this difference are not clear, except that the O_1 network may get confused by the orthographic information in cases where the vowel is not reduced.[1]

5.3 Word boundaries

Word boundaries seem to have an influence on loanword formation — especially in cases where many speakers of the borrowing language have at least some familiarity with the source language, in both its written and spoken forms (see section 2). This can most dramatically be seen in cases where an epenthetic vowel gets added to the end of the first word even though the sec-

[1]Another experiment was conducted in which the orthographic information was provided *only* for reduced vowels, but this network had an error rate of 12.8% (compared to 12.2% for the O_1 network).

ond word already began with a vowel — for example, *graphic-art* becomes /gurafikkuaato/ rather than /gurafikkaato/. We therefore made the simple modification of inserting an empty phoneme /_/ between words. As can be seen from Table 2, this leads to a reduction in error rate of anything from 0.6 to 1.8 percentage points for the neural networks, 1.1 to 1.9 percentage points for the decision trees.

5.4 Stress

Since stress has been reported to influence loanword formation in other settings (Silverman, 1992) it is natural to ask whether it might also play a role in the present context — for example, in determining gemination or vowel length. In order to test this, we added an extra input which indicates the stress of the current (or immediately preceding) vowel. A value of 1.0 was used to indicate primary stress, 0.5 for secondary stress and 0.0 for tertiary stress. We see that the performance of the decision trees is improved by 0.1 to 0.4 percentage points in the absence of word boundary information, but makes no noticeable improvement when word boundary information is present. The effect of making stress information available to the neural network is rather intriguing. From the limited data available in Table 2, it seems that the inclusion of stress information actually *degrades* the performance of the neural network when used on its own (from 13.5% to 14.8% in the case of the original (F) network) but *improves* the performance when used in combination with other modifications (from 12.1% to 11.5% in the case of the FO_2 network). We can only hypothesise that, in determining the final form of loanwords, stress interacts with other factors in subtle ways which are difficult for the neural network to capture reliably (especially in the absence of orthographic information), but are accessible to the more flexible structure of the decision trees. In future work, we hope to investigate this issue further, with a more thorough linguistic analysis.

5.5 Additional training

Networks FO_2B and FO_2BS were run for 100 epochs in order to compare them more closely, and to check for evidence of overtraining. Figure 2 shows the percentage of errors made by network FO_2BS for each of the three output groups. After 40 epochs the training and test errors, respectively, reach a level of 4.7% (resp. 7.2%) for the head phoneme, 0.7% (resp. 1.3%) for the added consonant, and 0.8% (resp. 1.1%) for the added vowel, making a combined error of 6.2% (resp. 9.6%). After this, the training error continues to fall while the test error levels off.[2] In contrast, the (test set) error rate for network FO_2B was found to level off at 10.3%, indicating that FO_2BS is the superior architecture.

6 Conclusion

We have shown that both feedforward neural networks (using backpropagation) and decision trees (using C4.5) may be trained to predict the phonological structure of English loanwords in Japanese using a featural input and a phonemic output. The error rate for the decision trees is lower than that of the neural networks by approximately two percentage points. Furthermore, providing information about the stress pattern, orthography of reduced vow-

[2]Note: the test error was computed at the end of each epoch, while the training error was computed during the epoch. Therefore the training error may exceed the test error in the first few epochs.

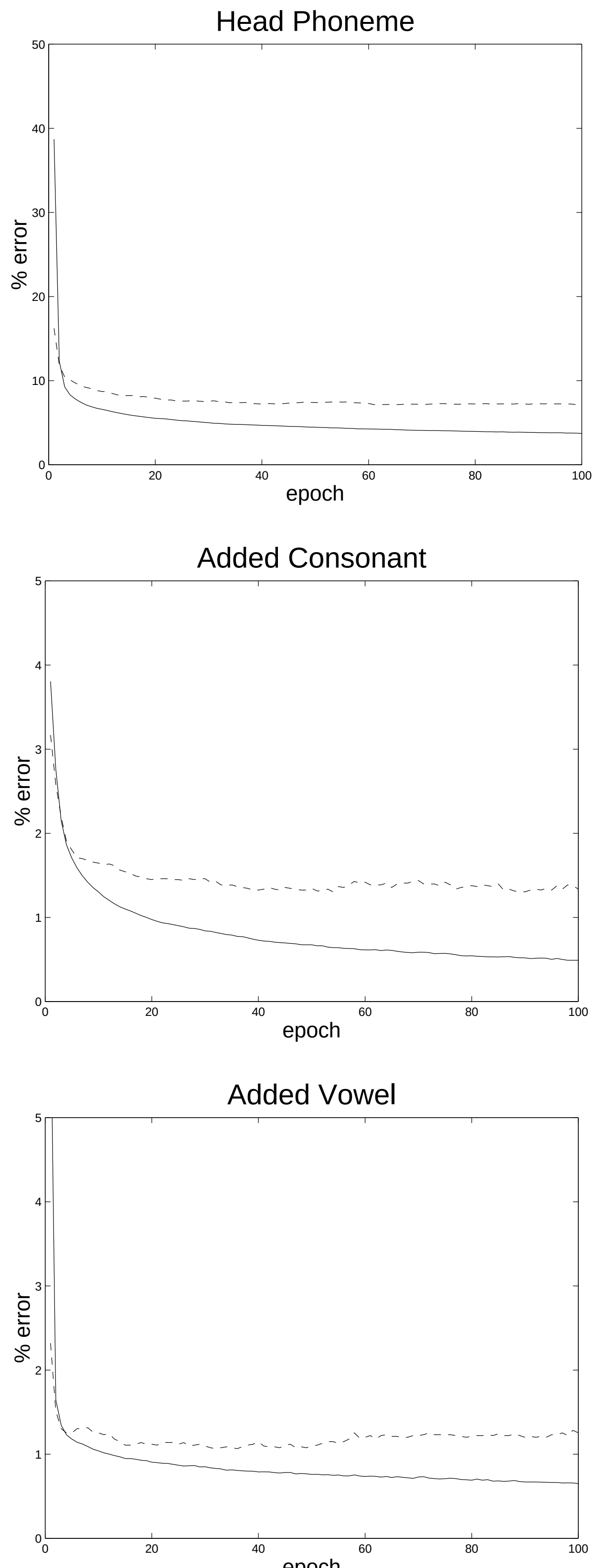

Figure 2. Percentage of training (solid) and test set (dotted) errors for the head phoneme, added consonant and added vowel.

els and location of word boundaries can reduce these error rates by a further four percentage points (from 13.5% to 9.6% for the neural networks, 11.5% to 7.6% for the decision trees).

These results are quite encouraging considering that, for the reasons discussed in section 2, the process of loanword formation has a certain element of indeterminacy in the sense that native speakers or dictionary compilers will often disagree on the correct rendition of a loanword in the orthography of the borrowing language. Many of the "errors" produced in our experiments were actually cases where a Japanese rendering has been produced which is different from that given in the database, but still an acceptable variant – for example, *glass* can be rendered either as /gurassu/ or /garassu/.

It seems that the decision tree learner can be relied upon to make productive use of additional information – whereas the neural networks appear on occasion to be confused by additional information, presumably due to the complexity of the interactions between different inputs. In particular, stress information seems to have a negative effect on the network's performance when provided on its own, but a positive effect when provided in combination with other factors.

In future work, we plan to make a more detailed examination of the two paradigms, taking into account the underlying strengths and weaknesses of each model and the types of errors produced, with the aim of introducing further modifications to reduce the error rate still further.

In addition, we aim to make a more detailed linguistic analysis of the network behaviour and the structure of the decision trees generated, in the hope that it may provide new insights into the underlying mechanisms of loanword formation.

References

Bailey, D.C., 1962. *A Glossary of Japanese Neologisms*, University of Arizona Press.

Blair, A.D. & J. Ingram, 1998. Loanword formation: a neural network approach, *Proceedings of the Fourth Meeting of the ACL Special Interest Group in Computational Phonology*, Montreal.

Gupta, P. & D.S. Touretzky, 1994. Connectionist Models and Linguistic Theory: Investigations of Stress Systems in Language, *Cognitive Science* **18**(1), 1–15.

Ingram, J.C., 1998. Word Prosody, lexical access and the perception of foreign loan words. In J.Wiles (Ed) *Perspectives on Cognitive Science II: Theories, experiments, and foundations*, Ablex.

Ito J. & R.A. Mester, 1996. Japanese Phonology, in J.A. Goldsmith (Ed.) *The Handbook of Phonological Theory*, Blackwell, Cambridge MA, 817–838.

Lovins, J.B., 1975. *Loanwords and the phonological structure of Japanese*, Bloomington : Indiana University Linguistics Club.

Quinlan, J.R., 1992. *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Rumelhart, D.E., G.E. Hinton & R.J. Williams, 1986. Learning representations by back-propagating errors, *Nature* **323**, 533–536.

Sejnowski, T.J. & C.R. Rosenberg, 1987. Parallel Networks that Learn to Pronounce English Text, *Complex Systems* **1**, 145–168.

Silverman, D. 1992. Multiple scansions in loanword phonology: evidence from Cantonese, *Phonology* **9**, 289–328.

Takagi, N. & V. Mann, 1994. A perceptual basis for the systematic phonological correspondences between Japanese load words and their English source words, *Journal of Phonetics* **22** (4), 343-356.

Appendix - Feature Matrix

	syll	cont	son	sib	nas	vce	lab	cor	high	back	low	aff	lat
p	0	0	0	0	0	0	1	0	0	0	0	0	0
b	0	0	0	0	0	1	1	0	0	0	0	0	0
t	0	0	0	0	0	0	0	1	0	0	0	0	0
d	0	0	0	0	0	1	0	1	0	0	0	0	0
k	0	0	0	0	0	0	0	0	1	1	0	0	0
g	0	0	0	0	0	1	0	0	1	1	0	0	0
C	0	0	0	1	0	0	0	1	1	0	0	1	0
J	0	0	0	1	0	1	0	1	1	0	0	1	0
f	0	1	0	0	0	0	1	0	0	0	0	0	0
v	0	1	0	0	0	1	1	0	0	0	0	0	0
θ	0	1	0	0	0	0	0	1	0	0	0	0	0
	0	1	0	0	0	1	0	1	0	0	0	0	0
s	0	1	0	1	0	0	0	1	0	0	0	0	0
z	0	1	0	1	0	1	0	1	0	0	0	0	0
	0	1	0	1	0	0	0	1	1	0	0	0	0
	0	1	0	1	0	1	0	1	1	0	0	0	0
h	0	1	0	0	0	0	0	0	0	0	0	0	0
m	0	1	1	0	1	1	1	0	0	0	0	0	0
n	0	1	1	0	1	1	0	1	0	0	0	0	0
N	0	1	1	0	1	1	0	0	1	1	0	0	0
l	0	1	1	0	0	1	0	1	0	0	0	0	1
r	0	1	1	0	0	1	0	1	0	0	0	0	0
w	0	1	1	0	0	1	1	0	1	1	0	0	0
y	0	1	1	0	0	1	0	0	1	0	0	0	0
i	1	1	1	0	0	1	0	0	1	0	0	0	0
e	1	1	1	0	0	1	0	0	0	0	0	0	0
æ	1	1	1	0	0	1	0	0	0	0	1	0	0
a	1	1	1	0	0	1	0	0	0	1	1	0	0
o	1	1	1	0	0	1	1	0	0	1	0	0	0
u	1	1	1	0	0	1	1	0	1	1	0	0	0
	1	1	1	0	0	1	1	0	0	0	0	0	0