EXPERIMENTS IN APPLYING MORPHOLOGICAL ANALYSIS IN SPEECH RECOGNITION AND THEIR COGNITIVE EXPLANATION

M A Huckvale Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, U.K.
A C Fang Department of Phonetics and Linguistics, University College London, Gower Street, London WC1E 6BT, U.K.

1. INTRODUCTION

1.1 Context

The majority of large-vocabulary continuous-speech recognition (LVCSR) systems for spoken English treat utterances as simple word sequences. The pronunciation of a sentence is predicted from the concatenation of the pronunciation of the words it contains. The probability of a sentence is predicted from the product of probabilities of the words in sequence (see e.g. [1]).

Not counting simplicity, there are many advantages to this approach: words are easy to isolate in text when statistical language models are being built; available pronunciation dictionaries are based on words; users expect to see the system recognise words; and the conventional measures of performance are based on words correct.

On the other hand, there is much evidence that human listeners are sensitive to the internal structure of words, particularly to that related to their meaning, or *morphology*. Some evidence for this is just to do with the production and interpretation of novel, morphologically complex words such as “un-micro-wave-ability”. But there is also interesting psycholinguistic evidence that morphological analysis is performed on-line during the processes of word recognition of common words. This seems to be the case for both inflexional morphology (*cows, milked*) and derivational morphology (*distrust, government*).

This paper contrasts some of the findings of these psycholinguistic experiments with our own experiments in applying morphological analysis within LVCSR.

1.2 Psycholinguistic findings

The psycholinguistic evidence for on-line morphological analysis stems from *priming* experiments. In these studies, the time taken for subjects to make a linguistic decision about a probe word is measured under different conditions. In the control condition an unrelated priming word is presented to the subject immediately prior to the probe word.

In the test condition, a priming word related to the probe word is presented. The experiments look for any systematic changes in reaction time as a function of the relationship between the priming word and the probe word. The findings reported below for morphological relationships come from [2].

What follows is a very simplified summary of the major findings. Readers are directed to the primary sources for more detail.

1. Similarity of phonological form is not sufficient to cause priming. For example *principal* does not prime *prince*. Even when there appears to be a morphological relationship, the meaning relationship must also be relatively obvious for priming to occur. For example *department* does not prime *depart*. Neither does priming occur for words with superficially obvious morphology when there is a perfectly good monomorphemic interpretation. For example *vanish* does not prime *van*. These results suggest that decomposition is not performed on the basis of the phonology alone.

2. Significant priming is observed between a stem and another word containing that stem combined with a suffix. For example, *attract* primes *attractive*, *attracting*, *attracted*. These results suggest that stems are linked to a family of morphologically-complex related words in the lexicon. Activation of the stem causes activation of morphologically related words.

3. Significant priming is observed between a word with a suffix and its stem. For example *attractive* primes *attract*. These results suggest that morphologically complex words are linked to their constituent stems in the lexicon. Activation of a morphologically complex word activates its stem.

4. Significant priming occurs even when the morphological relationship is obscured by complex morphophonological changes. For example *vain* primes *vanity*, and *teach* primes *taught*. These results seems to suggest that links between stems and their derived forms is not limited to simple phonological decomposition.

5. Priming is **not** seen between words with suffixes that share a common stem. For example *governor* does not prime *government*. This is in distinction to the fact that *govern* does prime *government*. These results suggests a time course to the priming phenomenon: that the priming effect occurs when the stem is heard, but that it is cancelled when more phonetic evidence comes in. That is, hearing *govern* primes both *governor* and *government*, but once the additional –*or* is heard, the priming on *government* is discarded.

6. Priming is seen between words with prefixes and words with suffixes sharing the same stem. For example *distrust* primes *trustful*. These results suggest that the priming effect can be handed on between word and stem and then between stem and word. There seems to be some asymmetry here with case 5, in that the presence of a prefix *dis-* does not preclude the interpretation *trustful*, even though the presence of a suffix *-or* precluded the interpretation *government*.

These are quite complex phenomena and quite difficult to account for in process terms. Cases 2 and 3 suggest that morphological analysis does take place during the recognition of some morphologically complex words. Case 1 shows that an additional restriction is that the morphology must reflect the meaning of the words too. Cases 1 and 4 seem to suggest that the decomposition is not simply phonological and we need to have some knowledge about the word before we attempt morphological decomposition. Cases 5 and 6 suggest that there are differences in the way prefixes and suffixes are processed.

1.3 Relationship to LVCSR

A speech recognition system working on whole words exhibits very few of the behaviours observed in the psycholinguistic data. The best that can be said is that such a system would exhibit cases 3 and 5. However this would be solely due to the degree of phonological overlap between the items. A system that is processing *attractive* might indeed add to the word lattice a hypothesis *attract*, while a system processing *governor* would be unlikely to add *government*. On the other hand, this system might also react to processing *principal* by adding a hypothesis *prince*, in contradiction to the findings in case 1.

Does any of this matter? Many engineers would argue that current LVCSR systems are just designed to maximise the probability of a signal given its interpretation; they are certainly not designed to mimic human beings. Many psycholinguists would argue that computational models that are not explicitly modelled on the workings of the brain can not tell you anything about cognition. This debate was reviewed in [3], which looked at the historical and philosophical basis for this mutual distrust. One important conclusion of this article was that engineering models and cognitive models are to some extent *complementary* rather than in conflict. Engineering models seek to reproduce human behaviour in the sense of the *primary* task of recognising words accurately. Cognitive models seek to explain the peripheral or *emergent* behaviour of recognition, such as response times. Cognitive models were not designed to provide high recognition accuracy, and engineering models were not designed to explain the results of psychology experiments.

We feel that there are advantages to examining the similarities and differences between the two scientific approaches to spoken word recognition. It may be that some emergent behaviours are common to man and machine and may be the property of any system simply trying to maximise its performance. For example, the bias to frequent words seen in many cognitive models is paralleled by the use of prior probabilities in machine models (see [3]). Conversely we feel that the dfferences between the emergent behaviour of the human and the machine might suggest ways to improve machine recognition. All that is required here is a belief that the human recognition system also tries to maximise accuracy. For example, findings in the semantic relationships between words (such as [4]) may suggest ways in which statistical language models could adapt to the topic of the sentence.

1.4 Morphological analysis in speech recognition

In the experiment described below, we put a layer of morphological processing within a conventional LVCSR system and report on the consequences in standard engineering terms: perplexities and word recognition accuracy. Our intentions were merely to see whether the incorporation of morphological processing would improve or worsen performance. We did not set out to replicate the psycholinguistic findings, which in any case we did not study in detail until after the experiment was concluded.

However, this has become a test case for the approach described in [3]. Does the incorporation of morphological analysis improve recognition performance? If so, one could argue that this is some evidence as to why humans appear to perform on-line morphological analysis in recognition. Does the system exhibit emergent behaviours similar to those found in the psycholinguistic data? If so, one could argue that this behaviour is simply a consequence of exploiting morphology and meaning to improve accuracy.

In section 2 we describe how morphological analysis was incorporated into the decoder, while in section 3 we report on the results of an experiment which compares a word-based and a morph-based recogniser on the same material using comparative metrics of performance. This shows that the morph approach has advantages in some circumstances. In Section 4 we show how a combined morph and word recognition system demonstrates an overall advantage to an all-word or all-morph system. In Section 5 we return to the psycholinguistic evidence reported above and make comparisons with the emergent behaviour of our system.

2. Morphological Analysis

To perform morphological analysis of words in a way that was compatible with the recognition decoder, we applied additional phonological constraints to the process of morphological analsysis. Phonologically-constrained morphological analysis (PCMA) is a decomposition of words into a sequence of prefix/stem/suffix morphemes constrained by both orthography and pronunciation [5]. This ensures that (i) each word can be mapped to a unique morph sequence, and vice versa; and (ii) that the pronunciation of each word is derivable from the concatenation of the pronunciation of the morph components.

Based on [6], a total of 115 prefixes and suffixes were built into the morphological analyser. Each of these was given one or more pronunciations, for example:

```
-NESS = n eh s, n ih s, n ax s
-ES   = s, z, ih z
RE#   = r eh, r ih
```

These pronunciations were then used by the morphological analyser to determine whether a candidate word can be decomposed into morphs subject to the constraint that the word pronunciation could be found among the possible morph pronunciation concatenations. As an example, the decomposition of abandoned into *abandon* + *-ed* is allowable because the pronunciation may be constructed from the parts:

```
ABANDON     = ax b ae n d ax n
-ED         = d
ABANDONED   = ax b ae n d ax n d
```

On the other hand, the decomposition of *academician* is not allowed since its pronunciation cannot be reconstructed from its parts, i.e., *academic* and *–ian*:

```
ACADEMIC    = ae k ax d eh m ih k
-IAN        = ia n
ACADEMICIAN = ax k ae d ax m ih sh n
```

A list of possible PCMA decompositions was generated using the available word types found in the British National Corpus (BNC; [7]) in combination with a British English pronunciation dictionary (BEEP, [8]). The dictionary contains 104,185 successful PCMA decompositions. Each entry has on average 2.3 morphological components. The dictionary may be obtained from the authors.

The PCMA dictionary allows the rapid translation of a word sequence to an acceptable morph sequence and vice versa. If a decomposition is successful, the word is represented as a sequence of its component parts with a trailing hash sign (#) indicating the presence of a prefix and a leading hyphen (-) indicating a suffix. As an example, *disregarded* is decomposed into three parts: *dis# regard -ed*, whose pronunciation is to be retrieved through concatenating the pronunciations for *dis#*, *regard*, and *-ed* listed in the pronunciation lexicon.

3. Exploratory Experiment

3.1 Word and Morph Dictionaries

Two sets of pronunciation dictionaries were created. The first set comprises three conventional dictionaries of fully inflected word forms. They contain respectively 20k, 40k, and 65k word types selected according to frequency of occurrence from 10 million words of the BNC training data. The pronunciations were found from BEEP, supplemented by pronunciations generated from a set of letter-to-sound rules developed by one of the authors.

Table I summarises the sizes and lists OOV rates for the three dictionaries when tested with one hundred sentences selected from reserved test data from the BNC.

TABLE I: A summary of pronunciation lexicons of inflected word forms with OOV rates on test set

Lexicon	*Size*	*OOV%*
word-20k	19,998	4.3
word-40k	39,994	2.5
word-65k	64,978	1.6

The second set comprised three pronunciation dictionaries that contain morphemes instead of word forms. They were created through the morphological analysis of the three conventional dictionaries of word forms. We thus obtained three morph-based dictionaries that covered at least the same words as the three word dictionaries. Table II summarises their sizes and OOV rates.

We see that PCMA analysis is capable of substantially reducing the size of pronunciation lexicons. The PCMA dictionaries are about 70% of the size of the fully word form dictionaries. Furthermore, this reduction in size is combined with a *decrease* in OOV rate. This is because the morph entries are able to generate more words than those covered by the equivalent full word form dictionary.

TABLE II: A summary of PCMA lexicons of morph units, showing reduction over word forms with OOV rates on test set

Lexicon	*Size*	*Red. (%)*	*OOV%*
morph-20k	13,370	33.2	2.8
morph-40k	25,158	37.1	1.6
morph-65k	46,000	29.2	1.0

3.2 Language Models

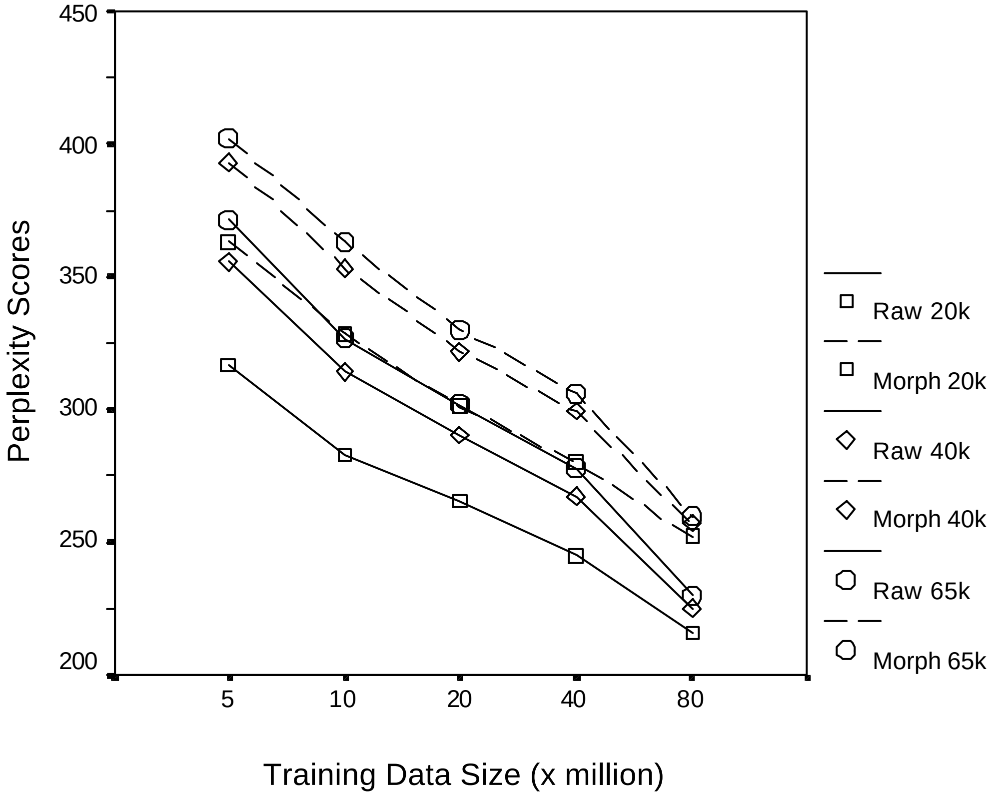
Subsets of the BNC were used as training data to build language models. The BNC data was first preprocessed to remove punctuation, and to convert numbers and abbreviations to words. Training subsets of 5M, 10M, 20M, 40M and 80M words were constructed. Each subset was also translated into morphs using the PCMA dictionary.

Word-trigram and morph-trigram language models were constructed using the CMU-Cambridge Toolkit [9] for each of the four sizes of training data for each of the six pronunciation dictionaries. Good-Turing discounting was applied to smooth the counts.

Perplexities were calculated on 100 sentences of reserved test data. The perplexity calculations are complicated by the fact that the test text increases in length after the mapping to a morph sequence. Since the definition of perplexity involves taking an average over the number of units in the text, we cannot simply run a perplexity calculation over the word and the morph sequences. Since there are more morphs, the average score will be less, even for the same total log probability. To make a better comparison, we find the total log probability for the test text and average over the number of stems rather than the number of units. As is usual practice, OOV words are not included in the calculation, although there are slight differences in OOV rates for the differerent models. Results are summarised in Figure 1.

Overall perplexities are about 10% worse for the morph models than for the word models. With 65k lexicons and 80M word of training data, the morph model gave a perplexity of 260, while the word model gave a perplexity of 230.

Figure 1 - Normalised perplexity for word and morph language models for different training data size



3.3 Word recognition results

For the recognition experiment, 100 sentences were randomly selected from the reserved portion of the BNC. These were read and recorded by a male speaker of southern British English in an anechoic environment. Finally, the recordings were digitally acquired at 16 KHz.

Recognition was undertaken using two software packages: word and morph lattices were generated from the speech signals with a demonstration version of the Abbot connectionist/HMM continuous speech recognition system [10]. Decoding of the lattices was performed by a decoder written at UCL.

Word and morph lattices were generated by Abbot using a set of parameters provided by Steve Renals (personal communication) which increased the number of word-hypotheses per node to a maximum of 100. Word lattices were generated using the language models trained on 80M words, while they were decoded using language models of different sizes.

Lattice scores

The lattice score is a measure of how many correct words from the test sets are present in the word lattices produced by the Abbot recogniser used for the experiments. To compute the lattice score, the path most similar to the correct answer was found through the lattice, and then the number of substitutions, deletions and insertions were counted on that path. Acoustic scores were disregarded. It is important to note that morphs could not be converted to word sequences in this measure, so that these are word accuracy for the word lattices, and morph accuracy for the morph lattices.

Table III shows how lattice scores vary across the choice of word and morph lexicons for the BNC test set, using language models trained on 80M words. The numbers in brackets are the percentages of completely correct sentences.

TABLE III: Lattice scores across the choice of word and morph lexicons for the LOB and BNC test sets, using language models trained on 80M words. Sentence inclusion rates in brackets.

Word Lexicon Size			*Morph Lexicon Size*		
20k	*40k*	*65k*	*20k*	*40k*	*65k*
86.6 (26)	88.6 (31)	93.0 (39)	93.1 (40)	94.7 (47)	95.5 (53)

The interpretation of these results needs to be performed with care, since although the word and morph lattices had the same depth, there are more words than morphs available to be put in the lattice. Nevertheless, we see no evidence from these results that the morph lattices are worse than the word lattices, and indeed argument could be made that they are superior. Certainly the morph accuracy is now close to 100% minus the OOV rate, and the sentence inclusion rates are considerably higher.

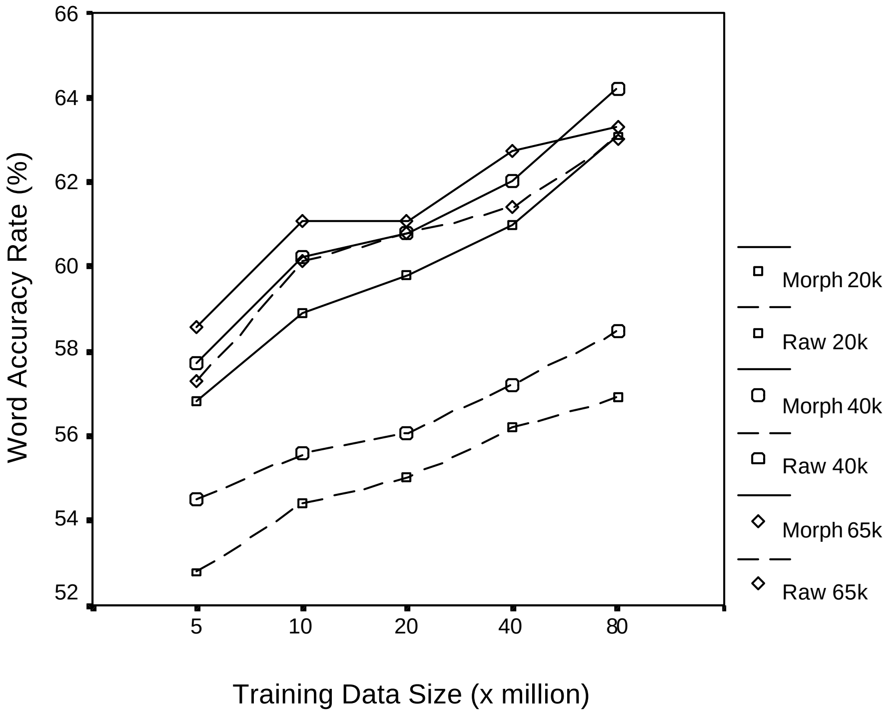
Word accuracy rates

To obtain recognition scores, the word lattices generated by Abbot were decoded with UCLdecode using the 30 different language model variants. Word accuracy was scored as before but on the best decoded path. Morph accuracy was calculated by concatenating affixes with their stems and then counting *words* correct, substituted or deleted. The results are shown in Figure 2.

From these results we can observe (i) that all systems improved in performance with increasing language model training data size; (ii) that all systems improved with increasing vocabulary size, although these improvements were less for the morph systems; (iii) the 20k-equivalent morph system consistently outperformed the 20k word system by about 4% absolute; (iv) the 40k-equivalent morph system consistently outperformed the 40k word system by about 3% absolute; (v) the 65k-equivalent morph and 65k word systems had very similar performance.

McNemar tests of significance using Peregoy's method showed that all differences between the morph and word results for the 20k and 40k lexicons were significant at the $p<0.01$ level; while there were no significant differences with the 65k lexicons.

Figure 2 - Word accuracy results for BNC test sentences
as a function of lexicon and language model
for word and morph approaches



3.4 Discussion

These exploratory experiments aimed to investigate the consequences of the shift from words to morphs in terms of lexicon size, OOV rate, perplexity, lattice score and word error rate.

In summary, we have seen that OOV rates can be reduced by about 1% absolute with a morph lexicon, with a simultaneous saving of 33% in lexicon size. Perplexity measures are worse, with perhaps an increase of 10% through the use of a morph trigram. Despite this, lattice scores and word recognition scores are similar, with morph systems even having slightly superior performance in some circumstances.

The biggest word accuracy gains are for configurations with the smaller vocabulary sizes, with the 20k equivalent and the 40k equivalent morph lexicons outperforming the 20k and 40k word lexicons for all quantities of training data. Performance increases with increasing training data size for all configurations, with only marginal evidence for a decrease in the benefit of the morph approach as the amount of language model training data increases. However there seems to be nothing to gain from the use of a 65k equivalent morph lexicon over a 65k word lexicon.

These exploratory results seem to show advantages for a morph system in circumstances where large lexica are impractical or unavailable, and that such gains are independent of the quantity of text available to language modelling (up to 80Mword at least). The fact that the performance of the morph systems holds up even in the worsening of the perplexity scores could well be due to the small improvements in the quality of the morph lattices over the word lattices.

4. Combination morph and word processing

The exploratory experiments showed little difference between the morph system and the word system for 65k vocabulary sizes. However they also showed that the perplexity of the morph language models was a litle worse, and the morph recognition lattices were a little better than the word models and lattices. This suggests that if we can combine the superior performance of the word language model with the superior morph lattice, we might obtain higher recognition performance. The challenge is to modify the decoder to input morph lattices but calculate probabilities using a word trigram language model.

To estimate the probability of attaching a new morph to a morph sequence with a word language model requires the calculation of these intermediate probabilities:

The probability of a morph given the morph sequence history is either

- $p_i^F = p(m_i^F \mid m_{i-N} \ldots m_{i-1})$ where m_i^F is the last morph in a word, and N is sufficient context to encompass three previous words
- $p_i^{NF} = p(m_i^{NF} \mid m_{i-N} \ldots m_{i-1})$ when m_i^{NF} may not be the last morph in a word, and N is sufficient context to encompass three previous words

To calculate these probabilities, it is necessary to chunk the hypothesised morph sequence into words, which is relatively easy to do using a morph dictionary. In the first case, the sequence forms exactly three words, and so the word trigram model can be used directly:

$$p_i^F = p(w_i \mid w_{i-2}\, w_{i-1})$$

However, in the second case, component m_i could be the prefix of many words, and the probability of this morph needs to be summed over all the possible matching words:

$$p_i^{NF} = \sum p(w_j \mid w_{i-2}\, w_{i-1})$$

where the sum is over all words w_j that match the end of morph sequence

Finally, the probability of the addition of a new morph m_i needs to be corrected for the assumptions made in the calculation of the previous morph in the hypothesis:

$$p_i = \begin{cases} \dfrac{p_i^{NF}\, p_{i-1}^{F}}{p_{i-1}^{NF}} & \text{if } m_i \text{ is start of new word} \\[2ex] \dfrac{p_i^{NF}}{p_{i-1}^{NF}} & \text{otherwise} \end{cases}$$

m_i starts a new word if m_{i-1} is a stem or a suffix and m_i is a stem or a prefix.

The synergy arising from combining morph lattices with word models is shown by improved word recognition accuracy. At the 80M and 65k level, the BNC test set gave an overall word accuracy of 66.3% for morphs with a word language model compared to 63.7% for words with a word language model and 64.5% for morphs with a morph language model.

5. Conclusions

The experiments described in sections 3 and 4 show that a very simple strategy for the incorporation of morphological analysis into a LVCSR system can lead to improvements in word accuracy. The findings of the exploratory experiment are that the productivity of morphs means that a morph lexicon of only 13,000 entries can outperform a word lexicon of 20,000 entries. However the findings are also that there is little benefit from this aspect of the analysis when the word vocabulary size reaches 65,000 entries. As far as the language modelling is concerned, the perplexity of a trigram morph model is much worse than the perplexity of a trigram word model. This is what one might expect in that a morph trigram uses less sentence context to base its probability estimate. Overall this first experiment does not provide convincing evidence for the utility of morphological analysis.

The findings of the second experiment with a morph lexicon and a word language model seem to show advantages that may arise from combining the best aspects of the morph and word approaches. The smaller number of lexical items seems to generate slightly superior morph lattices, which in combination with the word language model generates the highest word accuracy observed in this paper. It is interesting that it is the combination of the morphological analysis *and* the word sequence statistics that gives the best performance.

Returning to the psycholinguistic data presented in the introduction, we can now ask which of these observed behaviours are also shown by our recognition system. To do this we must make some assumptions about the relationship between psycholinguistic priming and some measurable property of our decoder. We will take a fairly broad stance in that we will assume that a probe word is primed if the word is present in the list of best hypothesis after the priming word is presented.

Let us take the data case by case, using the examples for clarity:

Case 1: (*vanish* does not prime *van*). Our system knows pronunciations for *van*, *-ish*, *vanish*. We can assume that all three will be hypothesised in the morph lattice. However since the word *van -ish* has never occurred it will be penalised by the statistical language model. Thus *van* makes it to the lattice, but not to the list of best hypotheses.

Case 2: (*attract* primes *attractive*). Our system knows the stem *attract* and it is put into the morph lattice. When the sentence fragment ending in *attract* is presented to the word model it must consider all words starting with the stem *attract* and so sums the likelihoods of each. In effect all words starting with *attract* are active in the list of best hypotheses.

Case 3: (*attractive* primes *attract*). Our system treats *attractive* as *attract -ive*. Both the stem *attract* and the suffix *-ive* make it through to the list of best hypotheses. The word interpretation *attract* is still present in the sense that it may later form part of another interpretation of the utterance. After all the sentence may be *attract eventually* and the *-ive* suffix incorrect.

Case 4: (*vain* primes *vanity*). Our system does not show this behaviour.

Case 5: (*governor* does not prime *government*). Our system treats *governor* as *govern -or* and *government* as *govern -ment*. The word *government* is considered when the input has reached the end of *govern* but when the final *-or* is processed, the properties of *government* are no longer relevant. It does not make it through to the list of best hypotheses.

Case 6: (*distrust* primes *trustful*). Our system treats *distrust* as *dis# trust*. If these two items are present in the word lattice then the statistical properties of *dis# trust -ful* will certainly be accessed in evaluating the sentence fragment. However the statistical properties of *trust -ful* will only be accessed if there is a competing explanation for the first morph, in for example the interpretation *this trust*.

Out of 6 cases, the system scores 4½. The most significant failure is to do with the more complex cases of morphological decomposition: those that involve significant phonological changes. Although one defence could be that such morphology is relatively rare in English, we have to admit that it is quite common in other languages, for example Arabic. Thus an interesting area for investigation would be whether a pronunciation dictionary which stored *vanity* as *vain -ity* and a language model which performed the same decomposition would lead to performance improvements. Or it may be necessary to build an explicit morphophonological processing layer in the decoder. This in turn might overlap with studies of phonological variation caused by accent variation and connected speech processes [5].

In this paper we have discussed an approach that we have taken to put morphology into speech recognition systems for English. We have shown that it is the combination of a morphological pronunciation dictionary in combination with a word language model which seems to offer improvements in performance. Whether by coincidence or otherwise, this combination of components also leads to emergent behaviour strikingly similar to that shown by human listeners in psycholinguistic experiments.

Acknowledgements

The work described in this paper was undertaken on a project "Enhanced Language Modelling" funded by the UK Engineering and Physical Sciences Research Council, Grant No GR/L81406.

References

[1] S. Young. Large Vocabulary Continuous Speech Recognition: A Review. In *IEEE Signal Processing Magazine,* 13(5), 1996, 45-57.

[2] W. Marslen-Wilson, L. Tyler, R. Waksler, L. Older. Morphology and meaning in the mental lexicon. In *Psychological Review*, 101, 1994, 3-33.

[3] M. Huckvale. Opportunities for re-convergence of engineering and cognitive accounts of spoken word recognition. In *Proc. Institute of Acoustics conference Speech and Hearing*, Windermere, 1998, 3-20.

[4] P. Zwitserlood. The Locus of the Effects of Sentential-Semantic Context in Spoken-Word Processing. In *Cognition,* 32, 1989, 25-64.

[5] A.C. Fang, M. Huckvale. Enhanced Language Modelling with Phonologically Constrained Morphological Analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2000, Istanbul, Turkey.

[6] R. Quirk, S. Greenbaum, G. Leech, J. Svartvik. *A Grammar of Contemporary English*, 1972, Longman, London.

[7] L. Burnard. *Users Reference Guide for the British National Corpus*, 1995, Oxford University Computing Services, Oxford.

[8] A. Robinson. BEEP Pronunciation Dictionary. Retrieved from World Wide Web: ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/beep.tar.gz

[9] P. Clarkson, R. Rosenfeld. Statistical Language Modeling using the CMU-Cambridge Toolkit. In *Proc. Eurospeech 97*, 1997.

[10] M. Hochberg, S. Renals, A. Robinson. ABBOT: The CUED Hybrid Connectionist-HMM Large Vocabulary Recognition System. *Proc. Language Technology Workshop, Austin Texas, Jan 1995*. Morgan Kaufmann.