Bootstrapping New Language ASR Capabilities: Achieving Best Letter-to-Sound Performance under Resource Constraints

Jim Talley

Human Interaction Research, Motorola Labs
9003 Bridgewood Trail; Austin, TX 78729; USA
Jim.Talley@motorola.com

Abstract

One of the most critical components in the process of building automatic speech recognition (ASR) capabilities for a new language is the lexicon, or pronouncing dictionary. For practical reasons, it is desirable to manually create only the minimal lexicon using available native-speaker phonetic expertise and, then, use the resulting seed lexicon for machine learning based induction of a high-quality letter-to-sound (L2S) model for generation of pronunciations for the remaining words of the language. This paper examines the viability of this scenario, specifically investigating three possible strategies for selection of lexemes (words) for manual transcription – choosing the most frequent lexemes of the language, choosing lexemes randomly, and selection of lexemes via an information theoretic diversity measure. The relative effectiveness of these three strategies is evaluated as a function of the number of lexemes to be transcribed to create a bootstrapping lexicon. Generally, the newly developed orthographic diversity based selection strategy outperforms the others for this scenario where a limited number of lexemes can be transcribed. The experiments also provide generally useful insight into expected L2S accuracy sacrifice as a function of decreasing training set size.

1. Introduction

One of the most critical components in the process of building automatic speech recognition (ASR) capabilities for a new language is the lexicon, or pronouncing dictionary. The lexicon is even a precursor to speech recordings to be used for training, since it is prudent to record phonetically balanced (or at least, phonetically diverse) materials when bootstrapping ASR in a new language, and the lexicon necessarily informs the development of such materials. However, in languages with few existing language resources, *e.g.*, Malayalam, it is unlikely that a quality electronic (or sometimes even, printed) pronunciation dictionary exists, and thus it must be created from scratch using available native-speaker phonetic expertise. Manual creation of an ASR lexicon is a difficult, labor-intensive process. It is, therefore, desirable to manually create only the minimal lexicon which will support machine learning based induction of a high-quality letter-to-sound (L2S) model which is capable of generating pronunciations for the remaining words of the language. The focus of this paper is on how to meet that goal – specifically, which strategy should be taken in selection of words to be manually transcribed for creation of a bootstrapping lexicon.

In brief, these experiments use an existing publicly-available, high-quality lexical resource (the CELEX [British] English lexicon) and state-of-the-art L2S model creation methodology (layered n-grams of multigram L2S correspondences) to test and contrast three strategies for selection of lexemes for training of a bootstrap L2S model. The three selection strategies – choosing the most frequent lexemes of the language (*Freq*), random selection of lexemes (*Rand*), and selection of lexemes using a newly developed orthographic diversity strategy (*Divers*) – are evaluated for both string (word) and phone generation accuracy on a strictly segregated test corpus over a range of bootstrapping lexicon sizes.

Details regarding the database, the modeling methodology, the word selection strategies, and the evaluation methods are included in section 2 below. Section 3 presents the results of the experiments and discusses them. And, finally, the findings are summarized and future extensions are discussed in section 4.

2. Structure of the Experiment

2.1. Database

This paper utilizes the well-known CELEX (British) English lexicon (Baayen, Piepenbrock, & Gulikers, 1995) as the database of pronunciations for training and testing purposes. Though this publicly-available pronouncing dictionary is cleaner and more consistent than most ASR lexical resources, our version does incorporate a number of modifications. Besides conversion into Motorola's fully IPA compliant ASCII representation, MAIPA (Melnar & Talley, 2003), the processing steps included: conversion of acronyms to period delimited sequences; conversion of all entries to lowercase only; separation of multiword expressions (with capture of novel pronunciation alternates); removal of several abbreviations (*e.g.*, "nom" [nominative], "comb" [combination], "arr" [arranged],…); identification and correction or removal of typos and errors; and, removal of some uncommon foreign words (*e.g.*, "tragedienne", "weltanschauung",…). After all of the clean up, slightly less than 44K lexical entries remained (lexical entry = lexeme + pronunciation).

2.2. L2S Modeling Method

The L2S modeling method was not itself the focus of this investigation, so a replication of the method of Bisani & Ney (2002) was employed. This state-of-the-art L2S modeling methodology layers a trigram statistical language model (SLM) on top of multigram L2S correspondences ("graphones" in their terminology). This method produces rather large L2S models but has the significant advantages of rapid training and quite good accuracy.

We already had high quality grapheme-to-phone correspondences (L2S mappings) in hand from other

internal L2S work, so it was not necessary to use the Bisani & Ney method of automatically producing those. Our L2S mappings allow 1-*n* letters to map to 0-*m* phones. The CMU-Cambridge SLM Toolkit (Clarkson & Rosenfeld, 1997) was used for SLM creation and probability estimation. It was driven by a custom beam search decoder for pronunciation predictions.

2.3. Word Selection Strategies

Three strategies for selection of a subset of the available lexemes are experimentally evaluated. Two (*Freq* and *Rand*) are reasonable default methods for choices of the items to have transcribed. The other (*Divers*) was specifically designed to address the problem discussed by this paper.

2.3.1. Most Frequent Lexemes (*Freq*) Strategy

The first strategy for word selection in our given scenario is the obvious possibility of simply selecting the most frequent words of the language for manual transcription. This strategy, which will be referred to as *Freq* below, is particularly appealing in that it has the side benefit of guaranteeing that the most common words of the language will have pronunciations provided by a human expert. This is important for two reasons: it is not unusual for irregularities (odd pronunciations, derivational exceptions, etc.) to be concentrated among the most common words of the language; and, human users of language technology are more accepting of mistakes by an automated system in places where humans also can be expected to have difficulties, and less accepting of the converse. For example, consider the likely effect on the user's impression of a speech synthesis system upon hearing the word "chimerically" mispronounced as / ʧ ɪ m ɛ ɹ ɪ k ə l ɪ / (substitution of / ʧ / for / k /) – if noticed, it would probably be dismissed as a reasonable mistake. In contrast, a typical user's response to hearing the same TTS system mispronounce the word "some" as / s oʷ m / might well be one of disgust – *i.e.*, finding it an incredibly stupid mistake that should not be made.

2.3.2. Random Selection (*Rand*) Strategy

Another obvious strategy, that of simply randomly choosing a subset of items from the full training lexicon, served as our second sub-selection strategy to examine. It is referred to as *Rand* in the text below. Given the stochastic nature of this strategy, four independent selections of 15K items were run. The resulting performance measures, appearing in the graphs and tables of section 3 below, reflect the mean accuracies for these four trials, with graphed error bars to characterize the *Rand* variance. (Note that the variance was so small that it was necessary to expand the error bars to three standard deviations in order to get the bars to even appear in the graphs.)

2.3.3. Diversity Based Selection (*Divers*) Strategy

And, finally, the third strategy which is examined is a newly developed strategy using information theoretic principles to select words based upon orthographic diversity. In this *Divers* strategy, the diversity measure is counter-balanced with word frequency in order to retain some of the desirable properties of the *Freq* strategy and avoid outliers (*e.g.*, misspellings, odd borrowings from other languages, etc.) in the master word list. The motivation for development of this strategy was the desire to maximize resulting models' predictive capabilities for the large number of words to be assigned pronunciations via L2S, given training on a relatively small sample of words. That is, we would like a model to be aware of, and take advantage of, the full range of orthographic-to-phonetic (or L2S) mappings which occur in the modeled language as a whole. But, of course, prior to creation of the lexicon, we do not have information regarding the range of such L2S mappings. Intuitively, however, novel orthographic sequences seem reasonable as approximate predictors of the desired novel L2S mappings.

In brief, this strategy takes a greedy approach to lexeme selection where a lexeme's fitness for selection, or score, is a weighted combination of three normalized factors: log frequency of occurrence, cross entropy of its sequence of orthographic characters, and string length of the lexeme. String length factors in very mildly as a tie-breaker / bias – all things being equal, we would rather have the longer word transcribed.

For a selection of N lexemes, we initialize the process by choosing $1/3 \cdot N$ lexemes based purely upon their frequencies of occurrence in the language. This initial "current set of selected words" serves to preserve, to some degree, a desirable property of the *Freq* strategy – having human expert-generated pronunciations for the important (and possibly anomalous) most common words. To proceed with picking of additional lexemes for inclusion, the current set of selected words is used to train a back-off trigram statistical language model (SLM) of letter sequences. That SLM is then used for letter probability estimation in cross-entropy calculations for ranking of the remaining candidate lexemes. Ideally, a new SLM would be generated each time a new lexeme is added to the selection set, but of course, that involves substantial computation with questionable real benefit. In practice, re-estimation of the SLM can be done periodically, after a number of lexemes have been chosen based on the cross-entropy estimates. For these experiments, a new SLM was generated after each 500 additional lexeme selections. The CMU-Cambridge SLM Toolkit (Clarkson & Rosenfeld, 1997) was used for SLM creation and probability estimation.

2.4. Evaluation Methodology

In this section, we cover the information necessary to understand what is reflected in the results presented in section 3. There is substantial discussion of issues in training and testing and methods of scoring.

2.4.1. Data Partitions

The full 44K lexical entry pronunciation dictionary was partitioned into mutually exclusive training plus *devtest* and *eval* testing subsets. The training subset consists of 80% (35K entries) and each of the two testing subsets contains approximately 10% (4.4K entries). Only the training and *devtest* partitions were used in these experiments. The partitioning scheme additionally guaranteed that all pronunciations for an individual lexeme were assigned to a single partition.

2.4.2. Stringent Pronunciation Scoring

With respect to multiple pronunciations, the accuracy measures reported herein are quite stringent. When we encounter multiple pronunciations (*i.e.*, more than one potentially correct answer), prediction is counted as wrong unless it matches exactly the single pronunciation of the lexical item which is currently selected for testing. It would be possible assume a lax scoring method – to count it correct if *any* of a lexeme's possible pronunciations matched (this would be more analogous to evaluations which use subjective human judgements). It is hard to say that one choice of how to score under conditions of multiple pronunciations is correct and the other is wrong. But, it is important to be aware of the two possibilities and their quite significant ramifications for reported scores. The stringent method of multiple pronunciation scoring used in these experiments guarantees scored errors in proportion to the number of alternate pronunciations. For example, with two pronunciations per lexeme, it would not be possible to achieve a Top1 string accuracy score better than 50% correct.

Related issues (with no clear resolutions) include arbitrariness (and (in-)consistency) in the choice of phonetic transcriptions and spotty inclusion of pronunciation variants across the lexicon as a whole. Given a list of L2S-transcribed lexemes which do not evaluate well under the metric used here, it has often been our experience that a linguist's review yields a response along the lines of "good… okay… that's fine… yeah, that's a possible pronunciation… pretty good…" The evaluation metric rigidly relies on the symbols present, whereas acceptable pronunciation is always on a "scale of grays." Among the factors that play a role in the spoken realization of a word are dialectal influences, formality, rate of speech, level of effort, and idiolectal peculiarities. Choices are, of necessity, made in the lexicon. For instance, take the English word "California". The end of the word might well be transcribed as /-nia/ or as /-nIa/. Of course, as rate of speech picks up (or level of effort decreases), the ending unambiguously becomes a single syllable, as in /-nya/ (which could also be fairly transcribed as /-ña/).

The point of the discussion in this sub-section is not to advocate the "correct" way to score, or to lament the choices that must be made in lexicon development, but rather to point out that 1) there are factors that significantly affect scores which one must be aware of in cross study result comparisons, and 2) given the lack of ability to capture the "gray scale" aspects of pronunciation, reported scores on the L2S task will likely be on the pessimistic side.

2.4.3. Scores Reported

We present, below, scores for both string (full lexeme [or word] pronunciation) accuracy and phone (individual symbol) accuracy. String accuracy is simply the number of test lexical item pronunciation strings which were exactly matched by the L2S predicted pronunciation – *i.e.*, the string correct rate. For the phone accuracy, we follow the spirit of the NIST scoring tools (Fiscus, 1995), using the Levenshtein Distance algorithm (see also http://en.wikipedia.org/wiki/Levenshtein_distance) to determine phone insertions, deletion, and substitution errors and, then, calculate phone accuracy as:

$$Accur = \frac{Corr - (Ins + Del + Sub)}{TotalTestPhones}$$

Because of the stringent multiple pronunciation scoring (as discussed in sub-section 2.4.2) and because of the intended ASR usage of the L2S generated pronunciations which easily admits multiple pronunciations, Top3 accuracy scores are presented below. Correctness within the best three generated pronunciations (especially phone accuracy) is seen as the most appropriate measure for evaluating L2S performance under the set of assumptions and the experimental structure.

It is recognized that string edit (Levenshtein) distance / phonetic symbol accuracy may not transparently map to ASR error rate (the measure which is ultimately of interest). Among the issues that make the mapping hard to predict are: 1) crazy pronunciation errors vs. plausible non-detrimental errors (*e.g.*, a (probably imperceptible) / ə / vs. / ʌ / confusion counts the same as a (glaring) / ɓ / vs. / s / confusion) – this is not quantified in current metric; 2) consistency may be more important than detailed accuracy (see, for example, Riley & Ljolje, 1996), especially when training and testing using a lexicon produced under matched conditions and modeling has ample representational power (as with HMM states with more than minimal Gaussian mixtures); and, 3) under ASR tests, words will have varied frequencies in the test corpus – i.e., not all lexeme pronunciations will have the same relative importance (in fact, the majority of words of the lexicon may not even be used in ASR testing).

2.4.4. Type vs. Token Frequency

An interesting issue in statistical L2S modeling is whether the statistics of lexical item types or tokens (or some combination of the two) should be used for model development and utilization, and whether type vs. token frequency should be taken into consideration in the evaluation metric. With type frequency alone, pronunciation predictions can be based up the simple frequencies of L2S correspondence occurrences in the lexicon – that is, if C("a"➔/e^y/)[1] is 1000 in the lexicon while Σ_i(C("a"➔P_i)) is 2500, then the probability of "a" mapping to /e^y/, p(/e^y/|"a"), can be coarsely estimated as 0.4. Token frequency, on the other hand, requires that relative lexical frequency be factored into the L2S correspondence probability estimation – *i.e.*, the token frequency of an L2S correspondence is the sum of its occurrences in the lexemes of the lexicon with each such lexeme occurrence multiplied the lexeme's frequency in a representative corpus.

Type frequency pays attention to the strength of an association pattern as indicated by the number of distinct lexical items which exhibit it. It is certainly plausible that speakers might be more likely to employ robustly instantiated patterns when pronouncing novel words or names. This would constitute a linguistic tendency toward regularization of the orthographic – phonetic mapping. Token frequency, on the other hand, gives considerably more weight to the patterns instantiated in common

[1] C("X"➔/Y/) is the count, or number of occurrences, of the orthographic characters of the string X mapping to the phone sequence Y in the training set.

words – almost ignoring possibly regular patterns which occur among rare words, preferring the familiar. The important thing to note here is that statistics based on type frequency vs. statistics based on token frequency make for quite different models.

Ultimately, it is a question for psycholinguistics to study and understand whether type vs. token frequency is the most appropriate model of human behavior in this domain. From casual observation, it is hard to say that either is clearly right or wrong. (My intuition is that humans, in effect, strike a compromise between the two.) Nonetheless, for all the work reported here only type frequency is considered. Note that switching to training and evaluating accounting for lexemes' token frequencies 1) would clearly boost the results for the *Freq* selection strategy, 2) would improve the results of the *Divers* strategy, but probably to a lesser degree than for *Freq*, and 3) would have an unknown (but quite possibly detrimental) effect on the *Rand* strategy's results.

2.4.5. Differences in Sampling Method

For these experiments, there was a difference in the way that lexemes were sampled for the *Freq* and *Divers* strategies, on the one hand, and the *Rand* strategy, on the other. For the former, samples were taken based upon unique lexemes. This was driven by the fact that the frequency information which plays into both the *Freq* and *Divers* strategies was available at the lexeme (*i.e.*, word) level, not at the lexical entry (*i.e.*, word + pronunciation) level. After selection of the lexeme, its first lexical entry (*i.e.*, first pronunciation) was selected from the training set for inclusion. Selection for the *Rand* strategy was performed on the full set of lexical entries (lexemes with pronunciations), rather than just taking first pronunciations for selected lexemes. This gave the *Rand* strategy (non-deterministic) access to the alternative (and first) pronunciations of those lexemes with multiple pronunciations. About 9% of the lexical entries in the full training set are alternate pronunciations.

3. Results and Discussion

The results of the experiments run are summarized graphically and numerically in the figures and tables below, where Figure 1 and Table 1 report on phone accuracy, and Figure 2 and Table 2 report on phone string accuracy (or the rate at which the full predicted phone strings exactly matched the corresponding "truth" pronunciation strings in the test lexicon). From the graphs, we can quickly note some general points.

Clearly, sub-sampling the available lexical items for training of a system for L2S does not perform as well as using the full available training set (97.6% phone / 87.9% string accuracy – dashed flat line towards the top of each figure) at least for the evaluated sub-samplings up to 15K lexemes. Note, however, that as the selection size increases the performance (of random selection in particular) is rapidly approaching that of the full training set. Projecting forward, we might expect accuracy to reach that of the full training set in the vicinity of 20K lexical entries (the full training set contains 35K lexical entries).

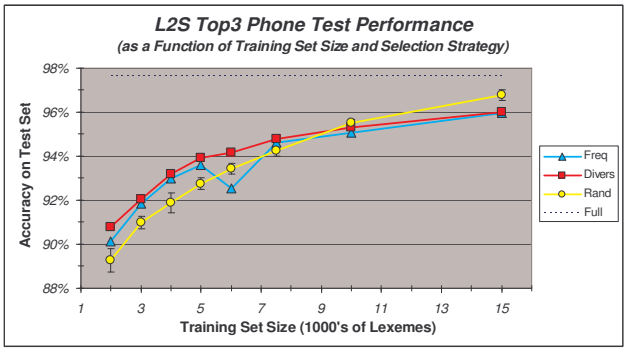

Figure 1: Graph of phone prediction accuracy on the test set under the various training set selection strategies as a function of the number of lexemes selected.

Letter-to-Sound Phone Accuracy (Top3)			
KLex	*Freq*	*Divers*	*Rand*
2.0	90.1%	90.8%	89.3%
3.0	91.8%	92.0%	91.0%
4.0	93.0%	93.2%	91.9%
5.0	93.6%	93.9%	92.7%
6.0	92.5%	94.2%	93.4%
7.5	94.6%	94.8%	94.2%
10.0	95.1%	95.3%	95.5%
15.0	96.0%	96.0%	96.8%

Table 1: Phone accuracies on the test set for various sizes of training sets when using the *Freq*, *Divers*, and *Rand* training set selection strategies.

The new *Divers* selection strategy clearly outperforms the *Freq* strategy, though not by a huge margin. This is true across the range of tested sizes, though there is convergence between results from the two strategies at the upper end of the training set sizes. The lexeme selection method used for these two strategies was the same, and we can, therefore, conclude that the *Divers* strategy should be preferred over the *Freq* strategy.

Note that around 6K lexeme selection size the *Freq* strategy exhibits an anomaly – accuracy falls dramatically. Further investigations (not presented here) with finer grained analysis around the anomaly showed the surprising result at 6K was not an error. Rather, those investigations made it clear that, despite the generally smooth trajectories of accuracy as a function of selection size apparent in the graphs, non-monotonicity (at least for the *Freq* strategy) is present.

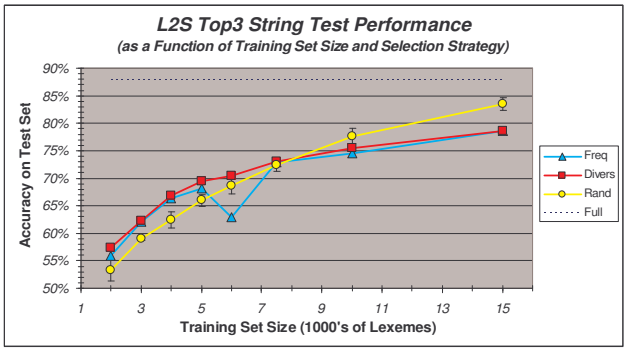

Figure 2: Graph of phone string prediction accuracy on the test set under the various training set selection strategies as a function of the number of lexemes selected.

Letter-to-Sound Word Accuracy (Top3)			
KLex	*Freq*	*Divers*	*Rand*
2.0	56.0%	57.3%	53.3%
3.0	62.0%	62.2%	59.0%
4.0	66.3%	66.8%	62.4%
5.0	68.2%	69.4%	65.9%
6.0	63.0%	70.4%	68.7%
7.5	72.8%	73.1%	72.4%
10.0	74.5%	75.5%	77.6%
15.0	78.6%	78.7%	83.4%

Table 2: String accuracies on the test set for various sizes of training sets when using the *Freq*, *Divers*, and *Rand* training set selection strategies.

The graphs also clearly indicate that, for smaller training set selection sizes, both *Divers* and *Freq* strategies produce superior results to those from the *Rand* strategy, by quite substantial margins. And, then, interestingly, in these experiments the *Rand* strategy gradually closes the performance gap and outperforms the other two strategies as the number of selected lexemes increases. *Rand* finally crosses over the *Divers* curve at about 9K lexemes for phone accuracy (Figure 1) and at about 8K lexemes for string accuracy (Figure 2). Variance in results for different random selections was (somewhat contrary to my expectation) not a significant factor – the error bars on the *Rand* curves in the figures represent three standard deviations (*i.e.*, we have a very high degree of confidence that the true result lies in the contained range).

Unfortunately, the conclusion that the *Divers* selection strategy is superior to the *Rand* selection strategy cannot be fairly drawn based on these experiments. As discussed in sub-section 2.4.5, there was an inadvertent discrepancy in the way that lexical items were drawn. This discrepancy in methods of selecting from multiple pronunciations may have significantly affected the results. The *Rand* strategy would, for any given training set size, likely have somewhat greater diversity of L2S mappings relative to the other two strategies. For small training set sizes, even the most fundamental L2S mappings are probably somewhat undersampled, and this greater diversity may have exacerbated the undersampling effects, leading to losses in test performance. Conversely, for larger training set sizes, where there was sufficient sampling of basic L2S mappings, the greater diversity may well have given the *Rand* strategy an evaluation advantage. If this conjecture is correct, we would expect the (observed) effect of diminished performance for smaller training set sizes and boosted performance for larger training set sizes. Thus, the pattern of *Rand*'s accuracy vs. the other strategies' accuracies is possibly an artifact of the differing methods of lexical entry selection for the training sets. This needs to be investigated in future work.

4. Summary

This paper has generally examined the task of building letter-to-sound (L2S) capabilities under resource constraints – constraints on the amount of effort to be expended in doing manual phonetic transcription. Building up a pronunciation dictionary is an important first step in bootstrapping ASR technology for a new language. In our scenario, we assume that the pronunciation dictionary will be created by using native phonetic expertise to manually transcribe a subset of the lexemes of the language, and an L2S system will then be trained from the resulting seed lexicon. We specifically examined the issues related to which strategy should be used in selection of the subset of lexemes and how the success of the strategies varied as a function of the number of lexemes selected.

Three strategies were taken into consideration for the experiments reported on here: 1) the strategy of choosing the most common words of the language (*Freq*); 2) the strategy of randomly choosing from among the available lexemes (*Rand*); and 3) a new strategy, presented in sub-section 2.3.3 above, which utilizes (among other factors) a measure of the lexemes' novelty (or diversity) with respect to the lexemes which have already been chosen for inclusion (*Divers*).

The *Freq* strategy has the highly desirable property of providing hand-generated pronunciations for the most common words of the language which will likely include the most significant of the words with irregular pronunciations. However, on well-matched tests, the *Freq* strategy is consistently outperformed by the new *Divers* strategy. Furthermore, the *Divers* strategy is designed to preserve (though to a lesser degree) the property of having pronunciations of the most important words of the language supplied by phonetic experts. So, *Divers* is clearly preferred over *Freq*.

Though, graphically, we generally see the *Divers* strategy outperforming the *Rand* strategy in these experiments, the relative desirability of *Divers* over *Rand* is not determined by these experiments for a couple of reasons. First, though we see a substantial advantage, in the experimental data, for *Divers* over *Rand* with smaller selected training sets, *Rand* substantially surpasses the performance of *Divers* with larger sets of training lexemes. And, second (and more importantly), these experiments were carried out with an infelicitous oversight regarding equivalence in the method of selection of lexemes' pronunciation for the *Rand* strategy vs. the other two strategies. This difference in pronunciation selection clearly has the potential to affect the results and, therefore, we must consider the results for the *Rand* strategy to not be comparable with those from the other two strategies.

For future work, of course, the obvious first step is to modify the pronunciation selection methods so that all of the strategies use an equivalent method, enabling fair comparison between all pairs of strategies. Given the scenario that drives this study (limited resources for hand transcription of lexeme pronunciations for a new language), the most realistic method would be, for all selections strategies, to choose from the lexeme set and then take all of the pronunciations of each lexeme for training, since a (manual) transcriber would, if appropriate, provide multiple pronunciations. However, this would likely lead to unequal training set sizes. To preserve the property of equal numbers of lexical items for training, the *Rand* strategy's selection method could be modified to choose only a lexeme's first pronunciation (as for the other methods in this experiment). However, in light of the good performance of *Rand* for larger seed dictionary selections, it might also be interesting to investigate the effect of using a method of random choice between a lexeme's pronunciations for all strategies (as effectively was done with the *Rand* strategy in this experiment).

The results of this experiment have suggested a possibly interesting modification of the *Divers* strategy. Such a modification would phase out frequency and diversity as primary factors in selection of lexemes as the number of lexemes selected rises, phasing in randomness in choices in its place. It is clear from the performance growth curves (Figures 1 and 2) that there is more to be gained from representative sampling (as for *Rand*) than from additional diversity (as for *Divers*) once sufficient training of the range of basic L2S correspondences has been achieved.

In this experiment, the quality metric is simply phone prediction accuracy (individually and as strings of phones). While this is a reasonable metric, a couple of modifications to the metric would make for interesting investigations. First, it would be nice for the metric to somehow take into consideration the relative importance of the words for which predictions are made – one way to do this would be to simply frequency weight the lexeme scores. A harder to implement, but desirable string match metric modification would be taking into account the relative egregiousness of the errors made. And, it would also be desirable to extend the study to consider the effect of these strategies on actual ASR system performance rather than just on the phone symbol prediction task.

5. References

Baayen, R.H., R. Piepenbrock, and L. Gulikers (1995). *The CELEX Lexical Database (Release 2)* [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]. http://www.kun.nl/celex.

Bisani, M. and H. Ney (2002). Investigations on joint-multigram models for grapheme-to-phoneme conversion. In *Proceedings of the 7th International Conference on Spoken Language Processing* (ICSLP-2002), Denver, Colorado, USA, pp. 105-108.

Clarkson, P.R. and R. Rosenfeld (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. In *Proceedings of the 5th European Conference on Speech Communication and Technology* (EUROSPEECH'97), Rhodes, Greece, pp. 2707-2710.

Fiscus, J. (1995). Speech Recognition Scoring Package (SCORE) Version 3.6.2. Found at http://www.nist.gov/speech/tools/index.htm.

Melnar, L. and J. Talley (2003). Phone Merger Specification for Multilingual ASR: The Motorola Polyphone Network. In *Proceedings of ICPhS 2003*, pp. 1337-1340.

Riley, M. and A. Ljolje (1996). Automatic generation of detailed pronunciation lexicons. In Lee, Soong, & Paliwal (eds.) *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer.