# Clustering-based Phonetic Projection in Mismatched Crowdsourcing Channels for Low-resourced ASR

**Wenda Chen** and **Mark Hasegawa-Johnson**
Beckman Institute
University of Illinois at Urbana-Champaign
USA
wchen113, jhasegaw@illinois.edu

**Nancy F. Chen**
Institute for Infocomm Research
A*STAR
Singapore
nfychen@i2r.a-star.edu.sg

**Preethi Jyothi**
Indian Institute of Technology Bombay
India
pjyothi@cse.iitb.ac.in

**Lav R. Varshney**
Beckman Institute
University of Illinois at Urbana-Champaign
USA
varshney@illinois.edu

## Abstract

Acquiring labeled speech for low-resource languages is a difficult task in the absence of native speakers of the language. One solution to this problem involves collecting speech transcriptions from crowd workers who are foreign or non-native speakers of a given target language. From these mismatched transcriptions, one can derive probabilistic phone transcriptions that are defined over the set of all target language phones using a noisy channel model. This paper extends prior work on deriving probabilistic transcriptions (PTs) from mismatched transcriptions by 1) modelling multilingual channels and 2) introducing a clustering-based phonetic mapping technique to improve the quality of PTs. Mismatched crowdsourcing for multilingual channels has certain properties of projection mapping, e.g., it can be interpreted as a clustering based on singular value decomposition of the segment alignments. To this end, we explore the use of distinctive feature weights, lexical tone confusions, and a two-step clustering algorithm to learn projections of phoneme segments from mismatched multilingual transcriber languages to the target language. We evaluate our techniques using mismatched transcriptions for Cantonese speech acquired from native English and Mandarin speakers. We observe a 5–9% relative reduction in phone error rate for the predicted Cantonese phone transcriptions using our proposed techniques compared with the previous PT method.

## 1 Introduction

Mismatched crowdsourcing is a recently developed method of acquiring transcribed speech in low-resourced and zero-resourced languages (Jyothi et al., 2016). It makes use of cross-lingual perceptions from speakers of high-resourced languages (e.g. English, Mandarin, etc.) when native speakers are unavailable for the target language. When an utterance is perceived by listeners or transcribers who do not speak the utterance language, they may misperceive its phonemes; we model this misperception as a noisy communication channel. The annotator's orthography from his or her native language will introduce further variations due to randomness in the phoneme to grapheme conversion.

The result of mismatched crowdsourcing is a set of transcriptions in, say, English or Mandarin annotation orthography. These mismatched transcripts are aligned, filtered, and decoded, using a maximum a posteriori (MAP) decoder, to compute a distribution over phone sequences in the target language (referred to as a probabilistic transcript or PT) (Hasegawa-Johnson et al., 2016). More accurate PTs could be derived by modeling crowd workers with different native backgrounds separately and merging their

*Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing*, pages 133–141, Osaka, Japan, December 11-17 2016.

cross-lingual misperceptions, after estimating how the phonemes in the transcriber languages can be mapped to the phonemes in the target language.

This paper provides a novel approach for merging cross-lingual perceptions from more than one language channel without using any *a priori* knowledge of phone mappings between the transcriber languages and the target language. Section 2 describes the dataset used in this work. Section 3 explores how phonetic and tonal confusions in mismatched transcriptions relate to the distinctive features of phonemes in the transcriber languages. Section 4 describes a two-step clustering technique for our transcription prediction task using a bipartite graph. Section 5 shows our experimental results on Cantonese speech using mismatched transcriptions in English and Pinyin from native speakers of English and Mandarin, respectively.

## 2 Data Preparation and Description

The original multilingual mismatched crowdsourcing corpus is described in (Chen et al., 2016). We use mismatched transcriptions from native speakers of English and Mandarin corresponding to roughly one-hour of speech in Cantonese. A total of 3443 short utterances in Cantonese were each transcribed in Pinyin by six Mandarin speakers and 8130 Cantonese utterances were transcribed in English (using nonsense syllables) by ten English speakers. Native phonetic transcriptions were available for 813 Cantonese utterances. Table 2 shows the phonetic transcription of a sample Cantonese utterance, along with pairs of English and Pinyin mismatched transcriptions corresponding to this utterance. The original corpus in (Chen et al., 2016) also consisted of Vietnamese speech data which is not used in this work.

| | | |
|---|---|---|
| **Cantonese (original with Babel Lexicon)** | pin3 geung1 gan1 jyu6 le1 | |
| **Cantonese transcribed in English** | Transcriber number #1 | hing kung gun chi |
| | Transcriber number #2 | kin kup gun che |
| **Cantonese transcribed in Mandarin** | Transcriber number #1 | pin3 geng2 gen1 ju3 le4 |
| | Transcriber number #2 | pin2 gong4 gen1 ju2 ne1 |

Table 1: *Sample utterance in Cantonese with mismatched transcriptions in English and Pinyin.*

| **Mandarin** | p | $p^h$ | | k | $k^h$ |
|---|---|---|---|---|---|
| **English** | | $p^h$ | b | | $k^h$ |
| **Cantonese** | p | $p^h$ | | k | $k^h$ |
| **Syllabic** | - | - | - | - | - |
| **Sonorant** | - | - | - | - | - |
| **Continuant** | - | - | - | - | - |
| **Labial** | + | + | + | - | - |

Table 2: *Example of Phoible Table for the Languages.*

Each annotator's error rate is estimated as the average string edit distance from his or her annotations to those of every other annotator using the same orthography as in (Jyothi et al., 2016). Between 2–6 Mandarin annotators and 2–6 English annotators with the lowest average pairwise string edit distance are selected for further analysis. (Section 5 compares results using 2, 3, or 6 annotators per annotation language). PTs are computed by aligning all the transcripts specific to a particular transcriber language i.e., the English and Mandarin transcripts are aligned separately to form two sets of PTs. The Mandarin

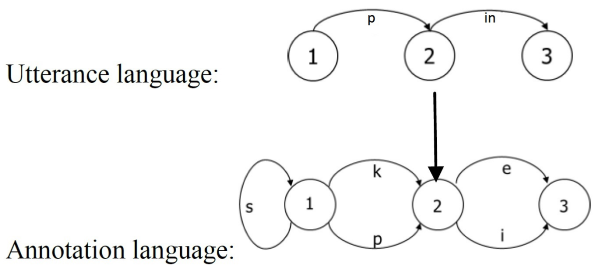

Figure 1: *FST Network Transfer*

PT and English PT are then each aligned with an utterance language transcript using a maximum likelihood alignment algorithm (Fig. 1 for the sample Cantonese sentence), in which the log probability of any given phoneme substitution is proportional to the Hamming distance between distinctive feature vectors corresponding to the two phonemes. Distinctive feature values for each phoneme are obtained from the Phoible phonological database (Moran et al., 2014); Table 2 shows four distinctive feature values corresponding to five different phonemes from the languages used in this work.

Suppose that $X = [x_1, \ldots, x_n]$ and $Y = [y_1, \ldots, y_n]$ are two phonemes whose distinctive features are $x_k \in \{0, 1\}$ and $y_k \in \{0, 1\}$ representing positive (1) and non-positive (0) distinctive feature values, respectively. The Hamming distance between these vectors is

$$D(X,Y) = \frac{1}{n}\sum_{f=1}^{n} |x_f - y_f|. \tag{1}$$

After the phonemes are aligned, they are converted to IPA symbols based on standard Mandarin and English orthography. The phone error rate (PER) derived from the phone alignments is hence computed as:

$$\text{PER} = 1 - \frac{\text{T}}{\text{M}}.$$

where $T$ is the number of correct phone mappings based on IPA and $M$ is the total number of the aligned phone mappings.

Phone error rates of these alignments had been reported in (Chen et al., 2016), where it is observed that Cantonese phone transcriptions recovered from Mandarin transcribers were much more accurate than those recovered from English transcribers.

## 3 Feature Weightings Analysis

Suppose we consider a weighted Hamming distance between phonemes (instead of an unweighted Hamming distance as shown in Equation 1):

$$\Delta(X,Y) = \sum_{f=1}^{n} G(f)|x_f - y_f|.$$

In order to define $\Delta(X, Y)$, it is necessary to choose some criterion for defining the feature weights $G(f)$. One such criterion, defined in (Nerbonne and Heeringa, 1997), is the information gain; we will not explore information gain further in this paper because it requires text in the utterance language. Mismatched crowdsourcing, however, provides us with an alternative measure of the distance between phonemes. Let t be a grapheme in the annotation language (English or Mandarin). Let $0 \leq S_X(t) \leq 1$ be the frequency with which utterance language phoneme $X$ is aligned with annotation-language grapheme $t$. Then the distance between phonemes $X$ and $Y$ can be measured by the total variation distance (TVD) between their grapheme alignment distributions (Varshney et al., 2016),

$$B(X,Y) = \frac{1}{2} \sum_t |S_X(t) - S_Y(t)|.$$

TVD is defined in the range $0 \leq B(X,Y) \leq 1$. The more similar two phonemes are (as perceived by annotators who speak a given language), the more often they will be transcribed using the same grapheme, therefore the smaller will be the TVD between them. A reasonable model is that the probability of confusion, $1 - B(X,Y)$, is the product of individual distinctive feature confusion terms of the form $\exp(-G(f)|x_f - y_f|)$, therefore

$$1 - B(X,Y) = \exp(-\Delta(X,Y)).$$

Using this model, the vector of weights *G(f)* is estimated as

$$G = \arg\min_G ||FG - B||_2^2,$$

where $F(XY,i) = |x_i - y_i|$ is a matrix with a row for every pair of phonemes, and a column for each distinctive feature.

| **Features** | **Information gain weights** |
|---|---|
| **Low** | 2.9750 |
| **Back** | 2.9210 |
| **Tense** | 2.5247 |
| **Front** | 2.8905 |
| **Syllabic** | 2.8878 |
| **Tone** | 2.8878 |
| **Round** | 2.6673 |
| **Labial** | 2.6570 |
| **High** | 2.1660 |

Table 3: *Feature weighting targets for Cantonese phones (Information gain)*

| **Features** | **Weights predicted from Mandarin Transcribers** |
|---|---|
| **Front** | 0.3407 |
| **Low** | 0.2293 |
| **Tone** | 0.1698 |
| **High** | 0.1678 |
| **Tense** | 0.1678 |
| **Syllabic** | 0.1334 |
| **Back** | 0.1087 |
| **Labial** | 0.1087 |
| **Round** | 0.1087 |

Table 4: *Weights prediction for Cantonese from Mandarin transcriptions*

Table 3 shows the theoretical information gain of the distinctive feature weights computed from the phone occurrence frequencies of Cantonese. Tables 4 and 5 show the estimated feature weights from the TVD approximation. The list of features and the order for English and Mandarin are similar especially for front and low features. This demonstrates that, given the transcription data, we obtained the relative order of the weightings of the distinctive features to be similar to the actually information gain and important of the features in characterising phones.

| Features | Weights predicted from<br>English Transcribers |
|---|---|
| **Front** | 0.3575 |
| **Low** | 0.1868 |
| **Tone** | 0.1216 |
| **High** | 0.0940 |
| **Tense** | 0.0940 |
| **Labial** | 0.0919 |
| **Round** | 0.0919 |
| **Back** | 0.0919 |
| **Syllabic** | 0.0630 |

Table 5: *Weights prediction for Cantonese from English transcriptions*

| Can. Tones | C1 | C2 | C3 | C4 | C5 | C6 |
|---|---|---|---|---|---|---|
| **C1** | 0 | 0.1668 | 0.0337 | 0.1946 | 0.1898 | 0.1221 |
| **C2** | | 0 | 0.1352 | 0.0536 | 0.0322 | 0.0446 |
| **C3** | | | 0 | 0.1656 | 0.1617 | 0.0984 |
| **C4** | | | | 0 | 0.0224 | 0.0724 |
| **C5** | | | | | 0 | 0.0676 |
| **C6** | | | | | | 0 |

Table 6: *Total variation distance (TVD) between pairs of Cantonese tones, based on their alignment with Mandarin mismatched transcripts. The smaller the TVD between two tones, the more likely they are to be confused in an MAP decoding of the mismatched transcript.*

Next we apply the TVD analysis to the Cantonese tones (C1–C6) and Mandarin tones (M1–M4) with the tonal features described in (Chen et al., 2016). Table 6 shows the TVD between pairs of Cantonese tones, based on their alignments with Mandarin mismatched transcripts. We observe that Mandarin annotators have trouble creating a Pinyin transcript that distinguishes the Cantonese high vs. mid level tones (C1 and C3), or that distinguishes the low rising tone (C5) from the mid-rising (C2) or low falling (C4) tones.

Table 7 lists the raw probabilities on which Table 6 is based: the probabilities $p(M_k|C_k)$ that Cantonese utterance tone $C_k$ is transcribed using Mandarin annotation tone $M_k$. We see that the Cantonese low falling (C4) and low rising (C5) tones are each most frequently annotated in Pinyin using the Mandarin low falling-rising tone (M3), whereas all three Cantonese level tones (C1, C3 and C6) are most frequently annotated by the Mandarin high level tone (M1).

| CanTone vs ManTone | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| **C1** | 0.568 | 0.105 | 0.270 | 0.055 |
| **C2** | 0.426 | 0.157 | 0.385 | 0.030 |
| **C3** | 0.562 | 0.104 | 0.304 | 0.029 |
| **C4** | 0.396 | 0.134 | 0.436 | 0.032 |
| **C5** | 0.400 | 0.151 | 0.413 | 0.033 |
| **C6** | 0.463 | 0.126 | 0.371 | 0.038 |

Table 7: *Mismatched crowdsourcing substitution probabilities $p(M_k|C_k)$ of Mandarin annotation tone $M_k$ given Cantonese utterance tone $C_k$.*

## 4 Phonetic Clustering Algorithm

This section describes how we infer phone mappings between the transcriber languages and the target language. Specifically we describe a phonetic projection framework and clustering criteria with random projections. The problem is formulated as a bipartite graph clustering problem followed by segment classification. The clusters correspond to the segment list of the target language represented using binary feature vectors. This is similar to classification but we allow some segments appearing in English and Mandarin to not be mapped to any target segment. The experiment is evaluated with Cantonese. As illustrated in Figure 2, the task is to cluster the phone mappings in the data from two multilingual transcriber channels to be the phone classes in the target language based on the similarity of the distinctive features in the clusters and in the segment.

Suppose that we have mismatched transcripts in Mandarin and English orthography, but we do not have native Cantonese phone transcripts. Additionally, let us assume that we do not know the Cantonese phone set. Since we can no longer compute the TVD between Cantonese phone types, we instead compute the TVD between Cantonese phone tokens. Take one of the two probabilistic transcripts (English, say) to define the number of Cantonese phone tokens per utterance. Align the other PT to it (the Mandarin one). The Mandarin PT has one or two orthographic symbols (or a deletion symbol) aligned to every segment of the English PT; thus for each segment $X$, its substitution probability mass function (pmf) $S_X(t)$ has up to two nonzero entries.

We first aggregate these probabilities over all instances of the same English orthographic symbol, so that $S_X(t)$ is the probability that English orthographic symbol $X$ is aligned to Mandarin Pinyin orthographic symbol $t$. We then build a matrix $W$ whose $(i, j)$th element, $w_{ij}$, is the probability that English orthographic symbol $i$ is aligned with Mandarin orthographic symbol $j$. In order to avoid losing tone information, we define the Mandarin orthography to be composed of Pinyin onsets and tone-annotated rhymes. Thus, the sequence $< hai3, ya2, you1, len1 >$ is decomposed into the 8 graphemes $< h, ai3, y, a2, y, ou1, l, en1 >$, which are aligned to the English orthographic sequence $< ch, an, h, eihn, n, uw, l, ah >$.

After constructing the matrix $W$, the next step involves merging the English (A's in full English segment set $E$) and Mandarin (B's in full Mandarin segment set $M$) clusters. We perform the following bipartite graph clustering using the normalized distances defined below.

Generally, the similarity between two sets A and B where $A \in E$ and $B \in M$can be defined as:

$$W(A, B) = \sum_{i\epsilon A, j\epsilon B} w_{ij}.$$

Hence the distance and normalised distance between two clusters set $A$ and $B$ can be computed using:

$$\begin{aligned} d(A, B) &= W(A, B^c) + W(A^c, B) \\ &= \sum_{i\epsilon A, j\epsilon B^c} w_{ij} + \sum_{i\epsilon A^c, j\epsilon B} w_{ij}. \end{aligned}$$

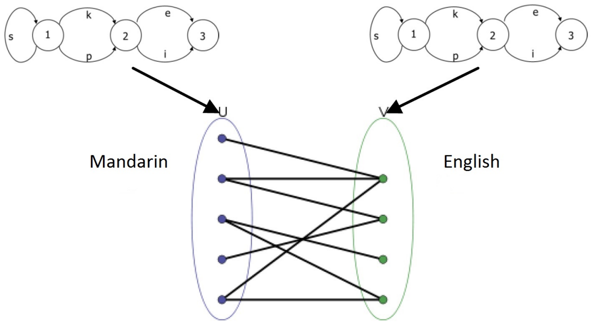

Figure 2: *From dynamic alignments to bipartite graph*

$$
\begin{aligned}
&d_N(A,B) = \\
&\qquad \frac{d(A,B)}{W(A,M)+W(E,B)} + \frac{d(A^c,B^c)}{W(A^c,M)+W(E,B^c)}.
\end{aligned}
$$

where $c$ is the conjugate sign of the set and the normalised distance is proposed to avoid the outliers in the set partioning. The final optimization criterion is then $\min_{\pi(A,B)} d_N(A,B)$ where $\pi(A,B)$ denotes partitioning into the $A$ and $B$ clusters.

The clustering algorithm is shown in (Zha et al., 2001) to be equivalent to the singular vector decomposition problem. The procedure is summarized in an algorithm called Spectral Recursive Embedding (SRE): given a weighted bipartite graph $G = (X, Y, E)$ with its edge weight matrix $W$ of the edge set $E$, we compute the scaled weight matrix and the second largest left and right singular vectors. Then we form partitions $A$ for vertex set $X$, and $B$ for vertex set $Y$ as the first cluster for the target segment. Subsequently we recursively partition the subgraphs $G(A,B)$ and $G(A^c,B^c)$ until we test and obtain the same number of clusters as the number of segments of the target language in Phoible.

In the two-step graph clustering process for Cantonese from English and Mandarin transcriptions, we 1) group the Mandarin tonal phones with different tones into 25 clusters, 2) group the aligned English phones into 25 clusters, and 3) finally, group the clusters on the two sides of the bipartite graph into 29 clusters. As analyzed in Section 3, the tones in the Mandarin transcriptions will be able to help the Cantonese transcription prediction process. Feature weights, estimated in Section 3, are also used in the clustering mapping and selection. The clusters are chosen and tagged with the target segments based on the largest number of common distinctive features. Feature weights are employed when two clusters could be tagged as the same target segment. For example, let us consider two clusters $A$ and $B$ that could be mapped to the same target segment $S$. Suppose segments in cluster $A$ are missing feature $F_1$ that appears in $S$ while segments in cluster $B$ are missing feature $F_2$ that also appears in $S$. If the feature weights determined for feature $F_1$ in Cantonese are less than the weights for feature $F_2$, then cluster $A$ is tagged to be the target segment $S$.

## 5 Experimental Results and Analysis

This section evaluates our clustering based method on Cantonese transcribed by 2–6 English-speaking transcribers and 2–6 Mandarin-speaking transcribers. For Cantonese, we have 1 hour of Cantonese speech accompanied by native transcriptions that can be used as our evaluation data.

Our method is evaluated by computing the most probable Cantonese phone sequence (including tones) given knowledge of the sequence of bipartite graph clusters. Let $X^\ell$ be the reference label of the $\ell$th Cantonese phone in a native transcription, and let $C^m$ be the cluster index, $1 \leq C^m \leq 29$, of the $m$th consecutive aligned set of Mandarin and English graphemes. The MAP Cantonese phone transcription is

$$
\begin{aligned}
&[\hat{X}^{(1)}, \ldots, \hat{X}^{(M)}] \\
&\qquad = \arg\max \prod_{m=1}^{M} p(X^{(m)}|X^{(m-1)}) p(C^{(m)}|X^{(m)}).
\end{aligned}
$$

where the language model $p(X^{(m)}|X^{(m-1)})$ is estimated from the grapheme-to-phoneme transduction of Cantonese text (Kong et al., 2016), and the misperception model $p(C^{(m)}|X^{(m)})$ is estimated using a separate training corpus with native and mismatched transcripts. The efficacy of the bipartite graph clustering algorithm could then be measured using the phone error rate (PER) between $\hat{X}$ and $X$. This could be compared with PERs of Cantonese transcripts recovered using only the English mismatched transcripts and with PERs of transcripts recovered using only Mandarin mismatched transcripts. All the target segments and transcription graphemes are converted into IPA phone set to compute phone error rates using the grapheme to phone conversion in (Hasegawa-Johnson, 2015). We also show PERs obtained using an FST union of the English and Mandarin mismatched transcript PTs, and from a majority

| **Cantonese Phone Error Rate** | $N$=2 | $N$=3 | $N$=6 |
|---|---|---|---|
| **Majority Vote** | 65.1% | 64.5% | 63.7% |
| **PT on English** | 64.3% | 63.2% | 62.7% |
| **PT on Mandarin** | 47.4% | 35.5% | 30.9% |
| **PT on E and M** | 43.1% | 30.6% | 29.5% |
| **Clustering method** | 39.1% | 25.5% | 27.9% |

Table 8: *Phone error rate (PER) for PT methods on Cantonese speech data. Here, $N$ corresponds to the number of mismatched transcriptions for each utterance.*

voting algorithm that outputs a symbol only if the Mandarin and English PTs agree. All the above-mentioned PERs are shown in Table 8. We found the optimal number of transcribers for two individual transcriber channels is 3 that helps compensate the language bias and variance across transcribers (i.e., noise in the mismatched transcriptions). The error rate of the clustering method slightly increases when more transcribers' alignments are combined, possibly due to higher variance across a larger number of mismatched transcripts. This can be improved by averaging and selecting the n best intra aligned transcriptions for clustering.

The distinctive features corresponding to each cluster combining English and Mandarin phones are a good match to the closest segment in Cantonese. Tone perception by Mandarin speakers provides some information about the segments of the target tonal language. The average number of edges per segment in the probabilistic transcription FST combining English and Mandarin transcriptions is 4.6. Although this threshold was carefully tuned on the evaluation data, the PTs combining Mandarin and English transcripts did not outperform the bipartite graph clustering algorithm.

The key comparisons that we note from Table 8 are: 1) PERs using the clustering method compared against the PERs from the English, Mandarin and English+Mandarin systems, and 2) PERs using the clustering method compared against the simple aligned majority voting method. We observe that the clustering method is significantly more accurate than the simple majority voting method which needs to use phone mapping knowledge between the target and transcriber languages. Our clustering method also improves over a system that uses only Mandarin mismatched transcriptions which indicates that we are able to leverage useful information from the English mismatched transcriptions. When a larger number of transcribers are available, despite the increase in variability in transcriptions, we observe that the clustering method is able to maintain good PERs by averaging the transcription alignments in the clustering process.

## 6 Conclusion and Future Work

This paper presents an extension of the mismatched crowdsourcing framework that makes use of mismatched channels corresponding to different transcriber languages. We propose a phoneme clustering algorithm that effectively combines mismatched transcripts from English and Mandarin native speakers to predict phone transcriptions for Cantonese speech. Future work includes applying the predicted transcriptions and projected segments in recognition tasks involving tonal languages.

## References

Wenda Chen, Mark Hasegawa-Johnson, and Nancy F Chen, 2016 "Mismatched Crowdsourcing based Language Perception for Under-resourced Languages, Procedia Computer Science, Volume 81, Pages 2329

Mark Hasegawa-Johnson, Preethi Jyothi, D. McCloy, M. Mirbagheri, G. di Liberto, A. Das, B. Ekin, C. Liu, V. Manohar, H. Tang, E.C. Lalor, N. Chen, P. Hager, T. Kekona, R. Sloan and A.K.C. Lee, 2016, ASR for Under-Resourced Languages from Probabilistic Transcription," in review

Preethi Jyothi and Mark Hasegawa-Johnson, 2016, Mismatched Crowdsourcing: A Novel Method for Acquiring Speech Transcriptions Using Non-Native Transcribers," in review.

Steven Moran, Daniel McCloy, and Richard Wright [eds]., PHOIBLE On Line. 2014.” Leipzig: Max Planck Institute for Evolutionary Anthropology (Available on line at http://phoible.org. Accessed on 2016-07-21)

Xiang Kong, Preethi Jyothi, and Mark Hasegawa-Johnson, 2016, Performance Improvement of Probabilistic Transcriptions with Language-specific Constraints. Procedia Computer Science 81:30-36

Lav R. Varshney, Preethi Jyothi, and Mark Hasegawa- Johnson, 2016, Language Coverage for Mismatched Crowdsourcing, Information Theory and Applications (ITA) Workshop, San Diego, California.

Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu, Bipartite Graph Partitioning and Data Clustering, 2001, Proceedings of the tenth international conference on information and knowledge management (CIKM 2001), pages 25-32

John Nerbonne and Wilbert Heeringa, Measuring Dialect Distance Phonetically, 1997, Proceedings of the Third Meeting of the ACL Special Interest Group in Computational Phonology

Mark Hasegawa-Johnson, 2015, SST Online Dictionary and G2P, http://www.isle.illinois.edu/sst/data/g2ps/