DECCA Repurposed: Detecting transcription inconsistencies without an orthographic standard

C. Anton Rytting and Julie Yelle
University of Maryland
Center for Advanced Study of Language (CASL)
College Park, MD
{crytting,jyelle}@umd.edu

Abstract

Most language resources and technologies depend on written text, while most endangered languages are primarily spoken. Transcribing speech into text is time consuming and error-prone. We propose a method for finding spelling inconsistencies without recourse to a standard reference dictionary or to a large training corpus, by repurposing a method developed for finding annotation errors. We apply this method to improve quality control of audio transcriptions, with particular focus on under-resourced, primarily oral language varieties, including endangered varieties.

1 Introduction

A critical part of documenting endangered languages is gathering and analyzing texts. In the case of many such languages, particularly ones without a long history of literacy or written literature, many if not most of these texts will be oral. Although recent work (Hanke & Bird, 2013) has explored ways of working with audio samples directly, most approaches to building additional resources (such as dictionaries and grammars, whether printed or digital) or human language technologies (such as part of speech taggers, morphological parsers, or automatic speech recognition systems) with audio text require transcription.

Even for languages with highly standardized spelling systems, maintaining transcription consistency is challenging. Inconsistencies in transcription can hamper the use of the corpus for other purposes, by distorting frequency counts and hiding patterns in the data. Transcription methodologies based on crowdsourced data collection have gained popularity in recent years due to their ability to deliver results at a fraction of the cost and turnaround time of conventional transcription methods (Marge, Banerjee, & Rudnicky, 2010) and collect linguistic data out of the reach of traditional methods of lexicography (Benjamin, 2015).

Yet crowdsourcing also carries a certain degree of risk stemming from the uncertainty inherent in the online marketplace (Saxton, Oh, & Kishore, 2013). While Marge, Banerjee, & Rudnicky (2010) found, for instance, that workers crowdsourced via Amazon Mechanical Turk (MTurk) had an average word error rate (WER) of less than 5% compared to in-house "gold-standard" transcription, Lee and Glass (2011) observed many MTurk transcriptions with a WER above 65%. Beyond concerns of authoritative knowledge and accuracy, the ability of crowdsourcing to open public lexicography to a "never-before-seen breadth of speaker input" from "the entire geographic range across which a language might vary" ushers in both insights and challenges related to language variation (Benjamin, 2015). More generally, any time the task of transcription extends beyond a small number of carefully trained transcribers, with limited resources for checking inter-transcriber agreement, the potential for inconsistencies arises.

We propose a simple, easy-to-apply method to examine transcriptions of audio corpora, including notes from elicitation sessions, for spelling errors and other inconsistencies that may arise in both conventional and crowdsourced data collection processes. While the proposed method is general enough to apply to any text input, we focus our experiments on transcriptions of spoken text. Our first set of experiments focus on transcriptions of spoken Arabic, including colloquial varieties; our second set focuses on the type of fieldwork we believe to be typical in building descriptions of (and resources for) endangered

Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages, pages 116–121,
Honolulu, Hawai'i, USA, March 6–7, 2017.

languages. In both cases traditional approaches to spelling correction do not apply, because there is no standard spelling dictionary to which to refer.

It is our hope that this method could assist field linguists in pinpointing aspects of transcriptions or other texts in need of quality control, reducing the need for manual examination of textual data.

2 Related Work

Much of the work on expediting transcription or providing quality control has focused on the needs of high resource languages. For example, Lee and Glass (2011) and Vashistha, Sethi, and Anderson (2017) assume access to an automatic speech recognition system in the language. Such methods will have little relevance for endangered language description, particularly at early stages.

So far as we are aware, there has been little work published on automatic methods for detecting inconsistencies in fieldwork or other transcriptions of spoken language without recourse to a standard lexicon or a large training corpus. However, there has been some work on two related problems: first, dealing with spelling variation in historical corpora (e.g., Baron & Rayson, 2008); second, detecting inconsistency of linguistic annotations such as part of speech (POS).

One approach to inconsistency detection in corpus annotation, called Detection of Errors and Correction in Corpus Annotation (DECCA)[1], postulates two root causes for variation—ambiguity and error—and posits that "the more similar the context of a variation, the more likely it is for the variation to be an error" (Dickinson & Meurers, 2003). Variation is defined as the assignment of more than one label (e.g., POS tag) to a particular word type (or, in the case of labels on phrases, a phrase type). Ambiguity occurs when more than one tag is appropriate for a given word or phrase type (e.g., multiple POS tags for an ambiguous word like "can"); an annotation error is an instance of a tag that is not appropriate for a token in context (e.g., a verb tag on "can" in the phrase "the can of tuna").

Intuitively, if a sequence of words is repeated multiple times in an annotated corpus, and a word within that sequence is tagged with different parts of speech in different instances of that sequence, it is likely that at least one of those tags is erroneous. Such a word sequence is called a *variation n-gram*.

3 Detecting Spelling Variants with DECCA

3.1 Defining the Task

In contrast to Dickinson and Meurers' interest in annotation errors, we are interested in detecting inconsistencies (or unwanted variation) in the text itself—the transcription of speech—without assuming the existence of any additional annotation layer such as POS.

Dickinson & Meurers' POS-tag error detection was performed in the context of a well-defined standard for annotation; thus, deviations from that standard may aptly be described as "errors." The tag set itself was a static, closed (and relatively small) set. In contrast, in our transcription-checking task, the list of words that may be used to transcribe a text is open and typically not predefined; even if dictionaries are used for guidance or reference, there may be words spoken (e.g., names, recent borrowings) that do not occur in the reference. For such words, at least, there may not be a pre-established standard spelling, and indeed for low-resourced languages there may be variant spellings for many words.

In this case, while finding spelling errors is still an issue, the larger question may be detecting variant spellings (such as "gr**ay**" vs. "gr**ey**" in English) that do not encode any semantic distinctions and hence are best conflated or unified to a single spelling for the purposes of (at least some types of) further analysis. Thus the detection either of a confirmed spelling error or of spelling variation like "gr**ay**" vs. "gr**ey**" that, in the judgment of a language expert, is not semantically meaningful (and hence can be conflated) would count as a hit for this task. Flagging a pair or set of words such as "pr**ay**" vs. "pr**ey**," which are legitimately distinct from each other, would be a false alarm.

3.2 Method

In order to apply DECCA to this purpose, we select some aspect of the speech transcription in which we suspect there may be inconsistencies—for example, whitespace or a particular spelling distinction—and we reformat the data such that that aspect of the transcription is removed from the context and treated as if it were a separate

[1] http://decca.osu.edu/software.php -- this work uses a modified version of the *decca-pos.py* program from DECCA 0.3, downloadable from RIDGES from the website https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/v4.1/decca-pos-reduce.py (as of February 15, 2017).

layer of annotation. If some inconsistency of that aspect is observed in contexts where the transcription is otherwise identical, the observed variation ought to be flagged for human review and possible normalization.

For whitespace variation, the presence or absence of whitespace characters (and other punctuation indicating word or morpheme boundaries, such as hyphens) within an otherwise identical phrase of at least two words is flagged for examination.

For detecting variation in spelling, our proposed method requires prior knowledge (whether from a language expert or some other source) of sets of substrings that may be sources of variation. For example, a language expert in English may flag the substring pair "**-ay**" vs. "**-ey**" as a potentially conflatable substring pair for finding hypothesized variants. In this case, the words "gr**ay**" and "gr**ey**" would be conflated in the text and the original spellings treated as tags in the DECCA input (e.g., {gr<1>, gray} and {gr<1>, grey} rather than Dickinson & Meurers' {word, POS} pairs such as {grey, ADJ}). Similarly, "pr**ay**" and "pr**ey**" would be mapped to {pr<1>, pray} and {pr<1>, prey}, respectively. DECCA then flags conflation terms for which multiple spellings occur at least once in the corpus in identical contexts (e.g., {"the gray dog"; "the grey dog"}). The intuition behind the heuristic is that true variants like "gray" and "grey" are more likely to show up in identical contexts than semantically distinct pairs like "pray" and "prey." Of particular interest are contexts with at least one word preceding and one word following the variant word—what we might call "non-fringe" contexts (following Dickinson & Meurer's heuristic of "distrusting the fringe").

4 Experiments

To test the feasibility of this approach, initial experiments were performed on corpora of transcribed spoken colloquial Arabic available from the Linguistic Data Consortium (LDC). We report here on two types of variation: spelling and whitespace.

After confirming the basic feasibility of this approach for spelling variation, we proceeded to test it on field notes on Kenyah Lebu' Kulit, an endangered language variety spoken in Indonesian Borneo.

Because we do not have complete ground truth on all the spelling inconsistencies for these corpora, we are not able to report recall. For these experiments, we therefore report precision only, based on an expert's review of the DECCA output in the various experiments.

4.1 Spoken Colloquial Arabic Transcripts

We tested DECCA's ability to detect inconsistencies in speech transcription on four corpora: GALE Phase 2 Arabic Broadcast Conversation Transcript Part 1 (Glenn, Lee, Strassel, & Kazuaki, 2013); Levantine Arabic Conversational Telephone Speech, Transcripts (Appen Pty Ltd, 2007); CALLHOME Egyptian Arabic Transcripts (Gadalla, et al., 1997); and CALLHOME Egyptian Arabic Transcripts Supplement (Consortium, 2002). These corpora provide transcripts of speech from three varieties of Arabic and at least five countries.

In these transcriptions, which reflect the diglossic nature of Arabic, orthography that is reflective of the colloquial pronunciation of dialectal words for which there are direct, nearly identical equivalents in Modern Standard Arabic (MSA) coexists alongside MSA orthography. As a result, these corpora, taken together, attest a considerable number of instances of spelling variation.

Two in-house Arabic-speaking researchers, one of whom is the second author of this paper, performed the human review of variation observed in identical contexts. Both annotators are native speakers of English who hold Master of Arts degrees in Arabic and certifications of Arabic proficiency at the ILR 3/ACTFL Superior level.

4.1.1 Spelling Variation

We summarize here experiments on two kinds of spelling variation in colloquial Arabic. One is the result of a phonological merger of two phonemes in some dialects (but not in MSA). The other variation in the spelling of the glottal stop, which in MSA is subject to complicated spelling rules, and in colloquial usage is often omitted. In the phoneme merger experiment, we examine the role of context frames in the precision of DECCA's hypotheses.

In the phoneme merger experiment, 63 pairs of words differing only in ذ (*dāl*) vs. د (*dhāl*), appearing a total of 242 times in the corpus, were flagged by DECCA. Each of these appeared at least once in the same context frame (non-fringe contexts, consisting of at least one preceding and following word: i.e., [*ContextWord1* _____ *ContextWord2*]). Of these 63 pairs, one of our in-house Arabic language experts judged that 61

were variant spellings that ought to be conflated, one was a semantically distinct minimal pair, and one was indeterminate. Excluding the uncertain pair, the precision was 98%.

Adding two other context frames—[*ContextWord1 ContextWord2* _____] and [_____ *ContextWord1 ContextWord2*] in addition to [*ContextWord1* _____ *ContextWord2*]—yielded an additional 68 items (appearing a total of 348 times), of which 60 were conflatable spelling variants, seven were semantically distinct, and one was uncertain. The combined precision was 94%.

If the context restriction is relaxed completely, then 337 items are returned, with 251 items consisting of variant spellings, and 80 records consisting of semantically distinct minimal pairs. Thus the baseline precision on the *dāl/dhāl* conflatable substring pair, without any restriction by context frame, is 76%.

For the *hamza* variation experiment, the second author of this paper annotated the 175 most frequent sets of words differing in *hamza* spelling that appeared in identical [*ContextWord1* _____ *ContextWord2*] frames. These sets of words appeared a total of 1,107 times in the corpus. Of these 175 items in context, two were semantically distinct minimal pairs, while 149 were variations that deserved normalization. Excluding 21 uncertain cases, the precision was 98.7%.

4.1.2 Whitespace Variation

In our preliminary whitespace experiments, we evaluated a subset of the variation instances with at least 20 characters of otherwise identical context, including at least two other consistent space characters, one on either side of the whitespace variation. The second author of this paper examined 95 variation n-grams, of which 22 were categorized as legitimate whitespace differences justified by semantic distinctions and 54 were seen as non-semantically motivated variants (31 errors; 23 instances of free variation). Nineteen items were marked as indeterminate and excluded. This yielded a precision of 71%.

4.2 Kenyah Lebu' Kulit

Having seen utility of this approach in our experiments on spoken Arabic transcriptions, we then applied the method to Kenyah Lebu' Kulit, an endangered language variety spoken by about 8,000 people in Indonesian Borneo. We used a database of transcribed texts available from the Max Planck Institute for Evolutionary Anthropology Jakarta Field Station as part of the "Languages of North Borneo" project (Soriente, 2015).[2] The data were inputted by a community-based documentation team, most of whom were native speakers of Kenyah Lebu Kulit without formal training as linguists. We were fortunate to obtain this corpus at an intermediate stage prior to the corpus collector's completion of quality control efforts leading up to publication, which allowed us to test our inconsistency-detection method's utility as an automated approach to quality control.

The Kenyah Lebu' Kulit corpus consists of 5,665 utterances in 27 files, with 52,549 tokens. This is considerably smaller than the corpus used in the Arabic experiments. Therefore we would expect fewer variation n-grams for any given experiment. Although we did not have access to an in-house expert in this language, the corpus contains Indonesian-language glosses that provided us with a rough indication of accuracy. We also consulted the researcher who collected the Kenyah Lebu' Kulit corpus, asking her to review instances of the corpus in which apparent orthographic minimal pairs appeared in identical contexts, to obtain verification that each pair was in fact semantically distinct (as described below).

As we examined this corpus, we noted that the orthography used by the transcribers draws a distinction between {é} /e/ and {e} /ə/. As experience working with informal texts in other languages has shown that diacritics such as the acute accent are frequently omitted, we first tested for inconsistencies between these two letters.

DECCA returned 31 orthographic minimal pairs differing only between {e} and {é}, five of which occurred in non-fringe contexts:

- <s> {mémang/memang} kaduq
- ngan {sénganak-senganak/senganak-senganak} teleu
- tei {é/e} </s>
- seken {mé'/me'} kena janan
- tegen {né/ne} ka senteng

Of these, the corpus collector confirmed that one ({é/e}) was a legitimate (semantically distinct) minimal pair, three were mistakes (two resolving to {…e…} and one to {…é…}), and the last ({né/ne}) was a mistake of a different sort: {né} should have been {né'} /neʔ/.

[2] Available from http://jakarta.shh.mpg.de/data.php (as of February 15, 2017). In these preliminary experiments, we used a version obtained from Dr. Antonia Soriente and Bradley Taylor.

Following up on a comment by the corpus collector about possible inconsistencies in transcribing word-final glottal stops, we conducted a further experiment examining the presence or absence of glottal stop at the end of a word.[3] As with the previous experiment, only non-fringe contexts were examined. This experiment yielded 102 variation n-grams, including 37 unique orthographic minimal pairs differing only by absence vs. presence of word-final glottal stop {'}.

Of these 37 orthographic minimal pairs, the corpus collector confirmed that only two ({lu/lu'} and {ra/ra'}) were semantically distinct minimal pairs. The other 35 were instances of spelling variation, yielding a precision of 95%.

5 Conclusion

Initial experiments suggest that DECCA can find inconsistencies in transcription of spoken Arabic, including both orthographic variation (assuming some prior knowledge of which sets of substrings should be examined) and variation in whitespace. Further preliminary experiments suggest that DECCA can also be applied to corpora collected in fieldwork settings, even when those corpora are relatively small.

We anticipate that DECCA, being simple and easy to use, could be applied as part of a suite of tools that field researchers and transcribers use for quality control on their own collections. We also anticipate it could be applicable for crowdsourced or community-led transcription efforts, particularly if wrapped in a user interface that facilitates the selection of candidate conflatable substring pairs and the reviewing of returned results. For example, the DICER tool (Baron, Rayson, & Archer, 2009; Baron & Rayson, 2009) could provide a framework for generating candidate conflatable substring pairs which could be input into DECCA.

Even for more resourced languages, where standard orthographies and reference dictionaries exist, this approach may prove helpful for words that may be missing from those dictionaries, such as names (particularly transliterated foreign names) and recent borrowings and neologisms. It may also help in instances where a standard dictionary includes multiple variants as equally correct, but greater consistency is desired by the corpus creators.

Although the focus of the work is on identifying spelling variants for quality control, insofar as the identification of minimal pairs may be useful for writing descriptive grammars and for training transcribers, we note that the tool can be used to identify minimal pairs as a byproduct of quality control.

Further work could focus on alternative context filters to improve coverage while maintaining high precision, particularly in the context of small corpora. We also welcome conversations with those who wish to apply this approach to corpora they are building.

[3] The orthography for Kenyah Lebu' Kulit uses the single quote mark {'} to indicate glottal stop. However, as word-final quote marks caused issues for the software used to manage the corpus, a {q} was substituted for most instances of word-final glottal stop in the intermediate stage of the corpus we worked with. For simplicity, we treat {'} and {q} as equivalent here, mapping {q} to {'}, anticipating a global mapping of {q} to {'} in published versions of the corpus.

Acknowledgments

The authors express their gratitude to Dr. Antonia Soriente and her funders for granting us access to the Kenyah Lebu' Kulit corpus. We thank Dr. Soriente for her generosity in evaluating our method's output, making it possible for us to test our transcription inconsistency-detection method on an endangered language. We also thank Valerie Novak for writing the scripts used to prepare the DECCA output of the Arabic corpora for annotation. Any errors remain our own.

This work was supported in part with funding from the United States Government and the University of Maryland, College Park. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government.

References

Appen Pty Ltd, S. A. (2007). Levantine Arabic Conversational Telephone Speech, Transcripts LDC2007T01. Web Download. Philadelphia: Linguistic Data Consortium.

Baron, A., & Rayson, P. (2008). VARD2: A tool for dealing with spelling variation in historical corpora. *Postgraduate conference in corpus linguistics.* Birmingham, England: Aston University.

Baron, A., & Rayson, P. (2009). Automatic standardisation of texts containing spelling variation: How much training data do you need? *Proceedings of the Corpus Linguistics Conference.* Lancaster, England: Lancaster University.

Baron, A., Rayson, P., & Archer, D. (2009). Automatic standardization of spelling for historical text mining. *Proceedings of Digital Humanities 2009.* College Park, MD: University of Maryland.

Benjamin, M. (2015). Crowdsourcing Microdata for Cost-Effective and Reliable Lexicography. In L. Li, J. McKeown, & L. Liu (Ed.), *Proceedings of AsiaLex*, (pp. 213-221). Hong Kong.

Dickinson, M., & Meurers, D. (2003). Detecting Errors in Part-of-Speech Annotation. *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03).* Budapest, Hungary: Association for Computational Linguistics.

Gadalla, H., Kilany, H., Arram, H., Yacoub, A., El-Habashi, A., Shalaby, A., . . . McLemore, C. (1997). CALLHOME Egyptian Arabic Transcripts LDC97T19. Web Download. Philadelphia: Linguistic Data Consortium.

Glenn, M., Lee, H., Strassel, S., & Kazuaki, M. (2013). GALE Phase 2 Arabic Broadcast Conversation Transcripts Part 1 LDC2013T04. Web Download. Philadelphia: Linguistic Data Consortium.

Hanke, F. R., & Bird, S. (2013). Large-Scale Text Collection for Unwritten Languages. *IJCNLP*, (pp. 1134-1138).

Lee, C.-y., & Glass, J. (2011). A Transcription Task for Crowdsourcing with Automatic Quality Control. *ISCA*, (pp. 3041-3044). Florence, Italy.

Linguistic Data Consortium. (2002). CALLHOME Egyptian Arabic Transcripts Supplement LDC2002T38. Web Download. Philadelphia: Linguistic Data Consortium.

Marge, M., Banerjee, S., & Rudnicky, A. (2010). Using the Amazon Mechanical Turk for transcription of spoken language. *IEEE-ICASSP*.

Saxton, G. D., Oh, O., & Kishore, R. (2013). Rules of Crowdsourcing: Models, Issues, and Systems of Control. *Information Systems Management, 30*(1), 2-20.

Soriente, A. (2015). Language Documentation in North Borneo Database: Kenyah, Penan Benalui and Punan Tubu. A joint project of the Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, the Center for Language and Culture Studies, Atma Jaya Catholic University and the University of Naples 'L'Orientale'.

Vashistha, A., Sethi, P., & Anderson, R. (2017). Respeak: A Voice-based, Crowd-powered Speech Transcription System. *CHI 2017.* Denver, CO: ACM.