# Tone restoration in transcribed Kammu: Decision-list word sense disambiguation for an unwritten language

*Marcus Uneson*

Centre for Languages and Literature, Lund University

marcus.uneson@ling.lu.se

## Abstract

The RWAAI (Repository and Workspace for Austroasiatic Intangible heritage) project aims at building a digital archive out of existing legacy data from the Austroasiatic language family. One aspect of the project is the preservation of analogue legacy data. In this context, we have at our hands a large number of mostly-phonemic transcriptions of narrative monologues, often with accompanying sound recordings, in the unwritten Kammu language of northern Laos. Some of the transcriptions, however, lack tone marks, which for a tonal language such as Kammu makes them substantially less useful. The problem of restoring tones can be recast as one of word sense disambiguation, or, more generally, lexical ambiguity resolution. We attack it by decision lists, along the lines of Yarowsky (1994), using the tone-marked part of the corpus (120kW) as training data. The performance ceiling of this corpus is uncertain: the stories were all annotated, primarily for human rather than machine consumption, by a single person during almost 40 years, with slowly emerging idiosyncratic conventions. Thus, both inter-annotator and intra-annotator agreement figures are unknown. Nevertheless, with the data from this one annotator as a gold standard, we improve from an already-high baseline accuracy of 95.7% to 97.2% (by 10-fold cross-validation).

Keywords: word sense disambiguation, Kammu, decision lists, lexical ambiguity resolution, tone restoration, legacy data.

# 1 Introduction

The RWAAI (Repository and Workspace for Austroasiatic Intangible heritage) project aims at building a digital archive out of existing legacy data from the Austroasiatic language family, not only linguistic but also encompassing general cultural heritage (musicological, anthropological, etc). This goal involves digitizing analogue data, converting existing digital formats into modern, non-proprietary if necessary, and providing machine-readable metadata descriptions.

A major part of the project's digitization efforts concerns a large collection on Kammu, an Austroasiatic language spoken in northern Laos. The collection, gathered from the early 70's and on, among other things spots what is for an unwritten language a very sizeable text corpus: a large number of mostly-phonemic transcriptions of spontaneous, narrative monologue, often with the original recording intact. A problem, however, is that four decades of data handling, even by a single person, will inevitably witness varying practices. For instance, there are different spelling conventions; preferably, these should be harmonized. More crucially, the early transcriptions, in contrast to later ones, do not have tones marked, which for a tonal language such as Kammu is essential (especially for computational applications). The present paper deals with inference of the missing tones by word sense disambiguation (WSD), similar to what Yarowsky (1994) applied to restoration of accents in Spanish and French.

The paper is organized as follows. Section 2 describes the data. Thus, we present Kammu in some detail: a few words on the typological properties of the language (2.1); on the work which has produced the data we use (2.2); and finally on the particular content of the transcriptions – namely folk tales in narrative, self-paced monologue (2.3). Section 3 presents the problem as an instance of WSD, the experimental setup, and the outlines of the algorithm; and motivates the choice of decision lists as machine learning approach. Section 4 comments the results and Section 5 concludes with a general discussion of the context of the task, including variant applications of the algorithm.

# 2 Background

## 2.1 On Kammu

The following summary is based on Svantesson (1983). The Kammu (or Khmu; ISO 693-3 `kjg`) is the largest minority language of Laos, spoken by half a million people in the northern regions of the country, as well as in adjacent parts of Vietnam, Thailand and China. It belongs to the Khmuic subgroup of the Austroasiatic languages.

Kammu is predominantly analytic, exhibiting no inflectional morphology. Derivational morphology is productive through reduplication, prefixation and infixation. New words are also coined by compounding. Kammu has no practical orthography; however, practically all Kammu speakers also know Lao, and the younger generation is literate in it.

Kammu is a tonal language. With only slight simplifications[1] every lexical entry is monosyllabic and can be assigned either high or low tone. At type level, the tones are about

[1] The simplifications are operationally motivated and phonologically rather than phonetically based. First, we assume that function words carry tone, which is true in a lexical sense but may not be obvious in connected speech (and the transcriber may or may not have marked tones in such cases). Second, we assume that all words are monosyllabic, whereas in reality there are many sesquisyllabic ("one-and-a-half-syllables") words, where a full tonic syllable is preceded by a reduced, "minor" one. In some cases, the minor syllable may carry tone, although

equally common. However, among the most common types (many of them function words), low tone is much more frequent (e.g. 19 out of the 20 most frequent types in our data), and thus low tone is significantly more common at token level (about 71% in our data). In a recent Kammu dictionary (Svantesson et al., in press), there are on the order of a thousand minimal word form pairs with respect to tone.

## 2.2 The Lindell-Raw Kammu collection

Most of the RWAAI Kammu material was collected by Kristina Lindell (1928-2005) and her most important consultant, transcriber, translator, and general assistant, the native Kammu Kam Raw (1938-2011). Kam Raw moved to Sweden in 1974, where he became a researcher in his own right (then usually known under his Thai name, Damrong Tayanin); he continued documentation of the Kammu language and culture until his death.

The Lindell-Raw collection is very large. It is also somewhat physically dispersed and, in the lamentable absence of its creators, confusingly organized. The collection is very varied with respect to content (linguistic, ethnographical, anthropological, musicological, botanical) as well as physical format (thousands of drawings, figures and photos; thousands of notes, translations, transcriptions written by hand, on typewriter, and on printers of the 80's; hundreds of reel-to-reel tapes and audio cassettes; hundreds of VHS cassettes and super-8 movies; hundreds of floppy disks and zip drives; hundreds of GB of content on portable drives). Before the start of the RWAAI project, only a small fraction of the data had been previously digitized, leaving plenty of excavation sites for the data archaeologist.

## 2.3 Kammu folk tales

From the point of view of a linguist, including the computational type, the most interesting part of the Lindell-Raw collection is perhaps the many sound recordings of folk tales, ranging from about 1 to 20 minutes in length. The conversational setting is monologic, spontaneous, narrative: the speakers know the general content, but the actual phrasing is mostly improvised. At least 700 such stories are known to have been recorded, by dozens of speakers (although it is not clear at this point that all of the recordings have survived). A sizeable subset (between 300 and 350 have been identified at the time of writing) was transcribed by Raw over the years. English translations of many of these, but by no means all, have been published in six volumes (Lindell et al., 1977, 1980, 1984, 1989, 1995, 1998), with a seventh, unpublished one close to being finished at the time of Raw's death. A small number of the transcriptions also have accompanying interlinear glossings (Figure 1).

As noted above, Kammu is an unwritten language. The transcriptions are in principle phonemic, using IPA, except for common use of capitalization and punctuation. However, as is natural, they also bear some hallmarks of an idiosyncratic, increasingly conventionalized orthography – for instance, when transcribing the Eastern Kammu dialect, where the tonal distinction is missing and instead corresponds to a contrast in voicing (Svantesson, 1989), Raw still marks tone as for the cognate of his native Northern variety.

We currently have a machine-readable corpus of around 175 stories thus transcribed, comprising ≈120kW in total (Section 3.2). This may be tiny for written languages, but it is very

it is usually predictable. Third, the parts of a compound each independently carry tone – here, we consider compounds as strings of independent, monosyllabic words.

| | | | | | | |
|---|---|---|---|---|---|---|
| Yèm | yə̀ | húal | snáa | kɔ̀ɔy | prɔ̀ɔm | yɔ̀ tèe. |
| Time | old | bear | those-two | squirrel | together | each-other |

Once upon a time the bear and the squirrel were close friends.

| | | | | | | |
|---|---|---|---|---|---|---|
| Yə̀h, | tnì | prɔ̀ɔm | yɔ̀ tèe | yɔ̀h | ə̀h | crɔ́, |
| Well, | there | together | each-other | go | make | weir, |

Well, there they went together to make weirs

| | | | | | |
|---|---|---|---|---|---|
| kɔ̀ɔy | lə̀ə | ə̀h | crɔ́ | òm | nɛ̀, |
| squirrel | then | make | weir | water | small, |

The squirrel built a weir in a small brook

| | | | | | |
|---|---|---|---|---|---|
| húal | lə̀ə | ə̀h | crɔ́ | òm | nám. |
| bear | then | make | weir | water | big. |

but the bear built his weir in a big river.

Figure 1: First lines of a (tone-marking) transcription of a recording of the Kammu folktale Lìaŋ húal káp kɔ̀ɔy "The story of the bear and the squirrel", with interlinear glossing.

sizeable as far as unwritten ones go. Although Raw originally intended this corpus as documentation, not as a computational resource, we have already used it (in conjunction with the Kammu dictionary mentioned above) for some basic computational applications, such as simple context-dependent spell checking and interlinear glossing.

Nevertheless, for documentation and computation alike, it is highly desirable to extend the corpus with the remaining transcriptions (after digitization). The main problem in so doing is that the early transcriptions, although well matched in terms of style and genre, are incomplete: they do not mark tones, which makes them significantly less useful for many practical purposes. In the following, we use methods for word sense disambiguation to supply the missing tones with good accuracy, substantially reducing the errors of an already-high baseline.

# 3 Experiment

## 3.1 Tone restoration through WSD

Word sense disambiguation (WSD) aims at automatically assigning the most appropriate meaning (or meanings, but usually only one is preferred) to a homonymous or polysemous word, given its context. Early work often targeted a lexical sample, with a small, predefined set of ambiguous words, like *interest* or *bank*. Later, the "all-words" task has become more common, where systems simultaneously disambiguate every single word of an input text, given the different possible senses for each (see Navigli (2009) for an extensive survey, or Agirre and Edmonds (2006) for a monograph). This standard problem phrasing is reasonably accurate also for tone restoration in Kammu, with a few notes:

**All-words task** Kammu tone restoration is more "all-words" than most: every single lexical entry is associated with either high or low tone; if tone is unmarked in transcription, the word form is usually ambiguous (for instance, in our 120kW corpus, about 70% of word tokens also occur with the other possible tone). Thus, the task also targets function words, which are of less interest to most WSD applications.

**Easy supervision and evaluation** As noted by Yarowsky (1994) for the similar case of accent restoration in French and Spanish, by stripping the accents (or tones, in our case) from existing transcriptions, we get a gold standard and a fully supervised task, with convenient and objective evaluation. By contrast, in traditional WSD, annotating new training data is expensive, and human inter-annotator agreement is relatively poor even for well-known resources.

**Word form required** For typological reasons, there is no reasonable way to predict the tone of a Kammu word from its context only. This is different from, say, the accent restoration case, where we sometimes may know the accent of a word without actually knowing the word itself (for instance, in French, we may for syntactic reasons be sure that an omitted word is a participle, and thus has acute accent).

**Minimal resources available** Compared to most WSD settings, the computational resources for Kammu are few (a small corpus and a dictionary; see next section) – no lemmatizers, thesauri, taggers, parsers, etc. Furthermore, as pointed out above, the few resources there are were produced for the purposes of language and cultural heritage documentation – if they also turn useful for computational applications, this is more of a fortunate but unintended side effect. Just as importantly, qualified transcribers and/or proofreaders of Kammu are very hard to find.[2]

## 3.2 Resources

The Kammu folktale corpus we have used in the experiments below comprises 119999 words, divided into 10526 sentences and 174 stories. We should note upfront that it does contain a significant number of transcription errors and, in particular, inconsistencies – the transcriptions were never intended primarily for machines. The relatively high error rate unfortunately makes the corpus less useful as a gold standard; among other things, the performance ceiling is unknown. Thus, even if we use the term "accuracy" below, the figures reported are rather agreement rates with respect to corpus annotations (just like they are for Yarowsky (1994)). That is, they can be compared to each other, but do not correspond perfectly to accuracy in an absolute sense and would very likely improve if the gold standard would contain no errors. Similarly, a phrase such as "35% error rate reduction" is really a lower bound – the true figure may well be much higher.

We prepared the corpus as follows. Case was normalized and sentences (or rather utterances) split up at punctuation end marks (`[:!.?…]`). We used 10-fold cross-validation: from the training data every tenth sentence, starting at sentence $1, 2, \ldots 10$ in the different runs, was removed and added to the test corpus. The reported results are the average of these 10 runs. Words with no tone or more than one (typically function words with tone unmarked, and compounds written without separator, respectively; about 3000 tokens,

[2] Jan-Olof Svantesson (p.c.) estimates that there are less than 20 people in the world with experience of transcribing Kammu.

or 2.5%) do not conform to either the target classifier's output, nor to current transcription conventions; these were excluded from classifications (although permitted, with tones stripped, in the contexts of other words).

## 3.3 A tonal classifier

As in most approaches to traditional supervised WSD, our classifier computes, represents, and stores the context for homographs with known classes. When encountering unclassified, ambiguous words, representations for these are similarly computed, and compared to the stored contexts of the training data. In our case, we based the classifier on *decision lists* (Rivest, 1987; Yarowsky, 1994, 1996). Very briefly, in decision lists, each feature-value pair implies a test (like a line in a case statement); as soon as a single test is true, the associated classification is output. We will use the term *rule* for the combination of a test and its associated classification. Crucially, at training time the rules are arranged, and at application time the associated tests are run, in (automatically estimated) descending order of *discrimination power* – a measure of how cleanly a feature assigns an ambiguous item into a single target class, given the item's context.

Decision lists are easy to implement, conceptually simple, and flexible. In particular, they readily accommodate different kinds of information, bypassing the difficult modelling of dependencies between heterogeneous feature sets. Also potentially irrelevant features can be specified, with little performance loss (except training time) – the relevant features float to the top of the ranking anyway.

To be sure, there are many more recent machine learning approaches to WSD, and any one of them would also have been a reasonable choice. However, although we will not pursue the issue very far in this paper, decision lists by design have an additional property which make them well suited for working with a noisy gold standard: every single decision can be attributed to a specific, human-interpretable feature-value pair. This property of "classification accountability" is useful when we wish to trace the reasoning behind a surprising (mis)classification – it may be due to an error in the target class of the training data; but it may also be caused by, say, a spelling error in the feature description. If we are interested in decreasing the noise of the gold standard, both possibilities are worth following up.

We implemented a decision list classifier similar to that of Yarowsky (1994) for binary homonym discrimination with main specifications as follows below (we refer to Yarowsky (1994) for the algorithmic details).[3] An excerpt of a decision list thus learned is given in Fig 1.

**Notation** We use $w$ for a word with some unspecified tone; $w^{\uparrow}$ ($w^{\downarrow}$ ) for one with known high (low) tone; $w^{?}$ for a word whose tone we wish to find out. We write $c(w)$ for the number of occurrences of $w$ in the entire corpus, and similarly $c(w, f_i)$ for the occurrences with feature $f_i$.

**Baseline** We used the most-common sense baseline: $w^{?} = w^{\uparrow}$ if $c(w^{\uparrow}) > c(w^{\downarrow})$, else $w^{\downarrow}$. This baseline is very high (>95%) and can be implemented in few lines of code (for a

[3] As a variant, we also applied *interpolated* decision lists to the problem, using the mechanism suggested in Yarowsky (1994). However, as we observed no performance gain whatsoever over ordinary uninterpolated decision lists, for reasons of space we restrict the following report to our experiments with the uninterpolated version, conceptually (and implementationally) much simpler.

| score | feature | value | output |
|---|---|---|---|
| 5.99 | coll+1 | pɨan | màan |
| 5.63 | coll-1 | ləə | màan |
| 5.08 | coll-1 | priaŋ | máan |
| 5.08 | bow | cəə | máan |
| 5.08 | coll-1 | ma | màan |
| 4.79 | coll-1 | kɔɔ | màan |
| ... | | | |

Table 1: Beginning of a decision list for the minimal pair máan ’to bury; to fade’/ màan ’pregnant; Burmese’ (irrelevant duplicate rules omitted). For instance, if the first test which returns true is the third (that is, maan neither has pɨan to the right (coll+1) nor ləə to the left (coll-1); but does indeed have priaŋ to the left), then classifier output is máan.

comparison, as noted above, the much cruder baseline of always choosing the globally most common low tone will only score about 70.5%). In addition, we can take max$(c(w^{\uparrow}), c(w^{\downarrow}))/c(w)$ as a basic measure of confidence in the baseline decision.

**Features** We used single-word fixed-width features up to $k$ = 2, for which we let ’coll-1’ denote the neighbour one step to the left (and similarly: ’coll-2’, ’coll+1’, ’coll+2’); and single-word non-positional context features (’bag-of-words’) up to window-size $n$; for which we write ’bow-n’ (i.e. $\{w|w \in w_{-n}, \ldots, w_{-1}, w_1, \ldots, w_n\}$). As a less off-the-shelf item, we encode a little bit of linguistic insight in a third feature ’onset’: certain onsets determine tone unequivocally (Svantesson, 1983). For instance, an empty onset (initial vowel) occurs only with low tone, whereas onsets /d, ˀw, ˀj/ occur only with high. Of course, for words which do occur in the training corpus, we have access to more reliable global frequencies and onsets will add nothing new; but they can be useful for unseen words.

**Discrimination power** We used the log-likelihood ratio with simple laplace (add-delta) smoothing ($\delta = 0.05$; e.g. Jurafsky and Martin, 2008, p. 134) applied to $P$:

$$\text{dp}(w, f_i) = \left| \log \left( \frac{P(w^{\uparrow}|f_i)}{P(w^{\downarrow}|f_i)} \right) \right|$$

# 4 Results

The results are given in Table 2. The decision list-based classifier significantly improves on the baseline (0.9722 vs 0.9572, $p \ll 0.001$, paired t-test), an error rate reduction of around 35%. We note that features coll+2 and coll-2 are hardly ever selected, and that the width of the bag-of-words window for our tiny corpus reaches maximum already at 3 words.

The log-likelihood discrimination scores can be used for other purposes than rule ranking. For every word in the test corpus, if present in the training data, some rule will be the first to match, and it will have an associated log-likelihood. We can treat this score as a measure of classifier confidence and rank all classifications accordingly. Figure 2 shows a graph of cumulative accuracy versus algorithm confidence, thus operationalized.

| bow-n↓ coll±k→ | 0 | 1 | 2 |
|---|---|---|---|
| 1 | 0.9572 | 0.9654 | 0.9623 |
| 2 | 0.9690 | 0.9719 | 0.9715 |
| 3 | 0.9700 | 0.9722 | 0.9705 |
| 4 | 0.9686 | 0.9705 | 0.9701 |
| 5 | 0.9670 | 0.9693 | 0.9691 |

Table 2: Decision list tone restoration, accuracy. 10-fold cross-validation. Baseline 0.9572

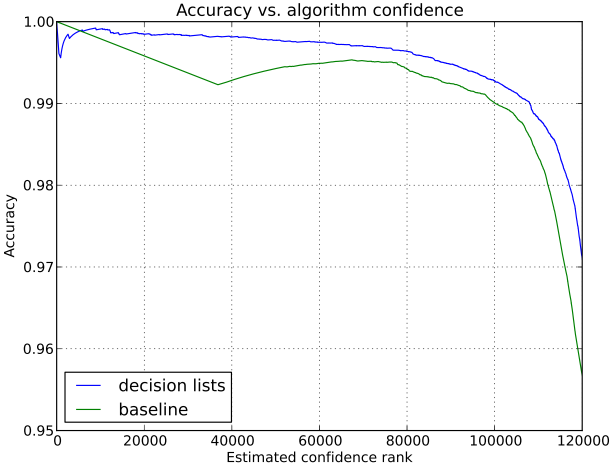

Figure 2: Cumulative accuracy per classification sites ranked after descending classifier confidence (summed results from all cross-validation runs). Decision lists (blue) and most-common baseline (green; see text for definition).

A strategy that lies close at hand is to apply the rules immediately and iteratively – conceptually (if not in implementation) only the most confident decision is applied in each iteration, and its result is allowed to influence future classifications. If the two members of a minimal pair with respect to tone occur in the same (tone-unmarked) context, then it is conceivable that classification accuracy for the pair may be improved if we first can assign tones to that context. As it turns out, this variation made no significant difference at all to our current setup, but we note the possibility for other tasks in the future.

## 5 Discussion

One might wonder if beating a good baseline with a percentage point or two is worth all the classifier hassle. It is, however. As noted, the Kammu corpus contains a certain amount of errors – typos, omissions, misinterpretations, competing conventions, etc – and qualified Kammu transcribers are in extremely short supply. Fortunately, a few of them are indeed tied to the RWAAI project. However, their time is valuable and needed for many things other than proofreading. In this context a single percentage point may mean 1000 errors less to worry about.

Actually, raw performance aside, the classifier may be even more useful from the more general perspective of managing human resources. If, as may well be the case, proofreading the entire corpus turns out infeasible, then an algorithm which can rank its decisions according to confidence may be used to point out where manual effort is likely to pay off best, in a process somewhat analogous to active learning (Settles, 2009). For instance, when checking automatically tone-annotated data, rather than traversing the corpus linearly, the transcriber/proofreader may check the nodes in ascending order of classifier confidence, as estimated by log-likelihood (cf. Figure 2). If desired, proofreading can be interrupted when the error rate falls below some threshold.

Similarly, it is easy to extract a ranked list over suspected mistakes in the training corpus: predict all tones by cross-validation, identify the miclassifications and sort them by descending order of confidence. With what we termed the accountability of decision lists (Section 3.3), this should be an efficient way of improving the gold standard, finding errors both in targets and in contexts.

On a final note, it should be pointed out that there is a recent Kammu-English dictionary (Svantesson et al., in press). We are currently converting it from presentationally oriented MS Word format into structured XML. Once this (rather painful) conversion is finished, the dictionary will certainly be an important resource. However, for the task at hand, its obvious usage (as a Kammu wordlist, disregarding the definitions) is less helpful than one might expect; its use is mainly restricted to word forms unseen in the corpus. For these, the wordlist can tell us whether they are unambiguous or not: in the former happy (but rare) case, it can output a classification; in the latter, it can at least provide better-coverage statistics on tones conditioned on word features such as onset. Also, for forms which are observed in the corpus with one tone only, the wordlist can possibly tell us that a minimal pair exists; this will not influence the system's guess, but may affect its confidence.

To be sure, there are more ambitious ways of using the dictionary, e.g. exploiting semantic overlap in the English definitions via Wordnet (Miller et al., 1990). Moreover, notwithstanding the desirable properties of decision lists (Section 3.3) for the task at hand, there are several other machine learning methods that could be tried. However, our next step should be to arrange proofreading of a large enough subset of the corpus to estimate the error rate of the gold standard. Any more ambitious attempts to extend the current system make little sense before this has been done, as it may well be the case that 97.2% is approaching the ceiling.

## 6 Acknowledgement

This work was carried out as part of the infrastructural project 'Digital Multimedia Archive of Austroasiatic Intangible Heritage', funded by the Riksbankens Jubileumsfond (project ID In11-0066:1). We are grateful to Jan-Olof Svantesson, Nicole Kruspe, and three anonymous reviewers for useful comments.

## References

Agirre, E. and Edmonds, P. (2006). *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science+ Business Media.

Jurafsky, D. and Martin, J. H. (2008). *An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice-Hall, 2 edition.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1977). *A Kammu story-listener's tales*. Number 33 in Scandinavian Institute of Asian Studies Monograph Series. Curzon Press, London.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1980). *Folk Tales from Kammu II: A Story-teller's Tales*, volume 40 of *Scandinavian Institute of Asian Studies Monograph Series*. Curzon Press, London.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1984). *Folk Tales from Kammu III: Pearls of Kammu Literature*. Number 51 in Scandinavian Institute of Asian Studies Monograph Series. Curzon Press, London.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1989). *Folk Tales from Kammu IV: A Master-Teller's Tales*. Number 56 in Scandinavian Institute of Asian Studies Monograph Series. Curzon Press, London.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1995). *Folk Tales from Kammu V: A Young Story-Teller's Tales*. Number 66 in Nordic Institute of Asian Studies Monograph series. Curzon Press, London.

Lindell, K., Öjvind Swahn, J., and Tayanin, D. (1998). *Folk Tales from Kammu VI: A Teller's Last Tales*, volume 77 of *Nordic Institute of Asian Studies Monograph series*. Curzon Press, London.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1990). Five papers on WordNet. *International Journal of Lexicography*, 3(4):235–244.

Navigli, R. (2009). Word sense disambiguation: a survey. *ACM Computing Surveys*, 41(2):1–69.

Rivest, R. (1987). Learning decision lists. *Machine learning*, 2(3):229–246.

Settles, B. (2009). Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Svantesson, J.-O. (1983). *Kammu Phonology and Morphology*. PhD thesis, Lund University. Travaux de l'Institut de linguistique de Lund, 18.

Svantesson, J.-O. (1989). Tonogenetic mechanisms in northern mon-khmer. *Phonetica*, 46(1-3):60–79.

Svantesson, J.-O., Tayanin, D., Lindell, K., and Lundström, H. (in press). Kammu yùan-english dictionary.

Yarowsky, D. (1994). Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 88–95. Association for Computational Linguistics.

Yarowsky, D. (1996). Homograph disambiguation in text-to-speech synthesis. pages 157–172.