CROSS-LANGUAGE PHRASE BOUNDARY DETECTION

Victor Soto[⋆] *Erica Cooper*[⋆] *Andrew Rosenberg*[†] *Julia Hirschberg*[⋆]

[⋆] Columbia University
[†] Queens College/CUNY

ABSTRACT

We describe models of prosodic phrasing trained on multiple languages to identify boundaries in an unseen language. Our goal is to create models from High Resource languages, in which hand-annotated prosodic phrase boundaries are available, to use in identifying boundaries in a Low Resource language, with little or no training material. We train models on American English, Italian, Mandarin, and German and test on each of these languages. We find that, while pause is the most important feature for phrase boundary prediction in all languages examined, the role of pause in boundary identification varies by annotator and the relative importance of other features varies significantly by language. We also find that different acoustic correlates of prosodic boundaries characterize different languages. In some, the relative importance of features is silence $>$ pitch $>$ intensity $>$ duration, while for other languages intensity is more important than pitch. These differences do not appear to be attributable to language family, since, e.g. English and German display different patterns.

Index Terms— Speech Understanding, Phrase Boundary, ToBI Breaks, Cross-Lingual

1. INTRODUCTION

Detecting prosodic events in speech has been shown to be useful for automatic corpus annotation for part-of-speech tagging, syntactic disambiguation, and text-to-speech corpora; for reducing language model perplexity for speech recognition; for salience detection; and for distinguishing between given and new information in speech summarization, identifying turn-taking behavior and dialogue acts in spoken dialogue systems [1, 2, 3, 4, 5, 6]. However, training prosodic event models typically requires a substantial amount of hand-labeled training data, which is not available for most languages. Previous attempts to train classifiers from speech for prosodic events such as phrase boundaries have relied upon such hand-labeled data with some success [7, 8, 9, 10, 11, 12, 13, 14]. However, there has been little work on phrase detection from classifiers trained on other languages.

In [15] the authors compare Mandarin and English break detection, though no cross-lingual validation is performed. They group intermediate and intonational phrase boundaries together and, using a set of acoustic, lexical and syntactic features, they perform within-language break detection. Our work here contrasts with this by presenting results on within-language, outside-language and leave-one-out experiments while using exclusively acoustic features. A similar approach to cross-lingual prominence detection on Standard American English (SAE), French, German and Italian was presented in [16]. These experiments showed that cross-lingual prominence detection is possible, although it also found important language-dependent differences and found little support for the hypothesis that language families might be useful for cross-language prosodic event identification.

Our current goal is to determine whether prosodic models trained on labeled data in one language can be adapted successfully to identify intonational phrase boundaries (IPBs) in another language for which little, if any, labeled data exists. We employ models trained on SAE, Mandarin Chinese, German, and Italian. In Section 2 we describe the corpora we use in our experiments. In Section 3 we describe our approach. In Section 4, we describe our cross-language adaptation and language-independent prediction experiments. In Section 5 we compare the important of different features for predicting phrasing in each language, and in Section 6 we describe out results in terms of the different characteristics of the languages examined.

2. MATERIAL

We examined four corpora, each representing a different language for our experiments: the Boston Directions Corpus, the DIRNDL Corpus, the Duration Corpus, and an Italian Read Speech Corpus. We note that the count of words does not include silences.

Boston Directions Corpus (BDC) – The BDC [17] is comprised of both read and spontaneous monologues elicited from four non-professional speakers, three male and one female. Speakers were asked to perform increasingly complex direction-giving tasks. Their directions were recorded and transcribed. Several weeks later, the subjects returned and were recorded reading transcriptions of their own directions. The corpus is orthographically transcribed and ToBI-labelled. We use three speakers for training material and the fourth (h2) as test material. The training split has 13,975 words (117 mins) with a phrasing rate of 16.65%; the test split contains 8,483 words (41 mins) with a phrasing rate of 14.82%.

DUR – The Duration corpus [18] is a dataset designed for the study of segment duration in Mandarin Chinese. The corpus was created using the ROCLING Chinese text corpus, from which 3845 phrases in total were extracted following a greedy algorithm that maximized coverage of relevant factors while minimizing the size of the resulting dataset. After the selection phase, the transcription and word segmentation of the text was manually corrected and recorded by a male native Beijing Mandarin speaker. The text was edited to

We thank Cinzia Avesani, Chilin Shih and Katrin Schweitzer for providing labeled corpora. This research was partially supported by 'la Caixa' Fellowship Grant for Post-Graduate Studies, Caixa d'Estalvis i Pensions de Barcelona, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Laboratory (DoD / ARL) contract number W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government.

match the recordings, and phrasing and prominence levels were also annotated. The dataset is split into training and test partitions with 75% and 25% of the original dataset respectively. The training split contains a total of 5665 words (78 min) with a phrasing boundary rate of 36.37% while the testing dataset contains 1964 words (28 min) with a phrasing rate of 39%.

DIRNDL – The Discourse Information Radio News Database for Linguistic analysis corpus [19] is a database of German radio news broadcasts. It contains approximately two and a half hours of radio news, along with accompanying transcripts from which fillers, disfluencies and music have been removed. We divide the material into training and testing splits with no speaker overlap. The training data has 12017 words (108 min) with an accent rate of 11.47%; the test data has 4323 words (40 mins) with phrasing rate 12.03%. The corpus is annotated for intonation according to GToBI [20].

Italian – Our Italian corpus contains about 35 minutes of read speech from a single male professional speaker; this corpus was made available to us by Cinzia Avesani at the Institute of Cognitive Sciences and Technologies in Padova. The speaker reads two different short stories. The corpus is orthographically transcribed and prosodically annotated for Italian ToBI. Since it contains material from a single speaker, the Italian experiments are speaker-*dependent* in contrast to the other corpora. The training data contains 2,756 words (26 mins) with a phrasing rate of 10.81%; the test data contains 1,166 words (9 mins) with a phrasing rate of 11.23%. We found that, in this corpus, boundaries marked as ToBI level 3 were similar in acoustic characteristics (particularly presence of pause) to boundaries marked as ToBI level 4 in our other corpora.

We begin our analysis by examining differences among the corpora with respect to the relationship of ToBI break indices of level 3 and 4 and the presence of pause in Table 1, and of overall IPBs and pause in Table 2.

	Break Labels			Mean Pause Dur (s)	
Corpus	3	4	Break 3-Sil	After 3	After 4
BDC	7.83%	15.96%	21.26%	0.1881	0.5092
DIRNDL	12.99%	11.61%	8.66%	0.1031	0.9236
DUR	13.07%	37.10%	25.58%	0.083	0.7397
Italian	9.41%	10.94%	95.66%	0.3565	0.7044

Table 1. ToBI break indices 3 and 4 summary

Corpus	IPB %	%IPB-Silence	% Silence	% Sil with IPB
BDC	15.96%	73.48%	14.72%	82.09%
DIRNDL	11.62%	93.84%	13.05%	82.54%
DUR	37.10%	61.02%	21.65%	83.70%
Italian	10.94%	98.93%	20.32%	42.08%

Table 2. IPB-related statistics for each full corpus.

This table shows the number of these that coincide with silence. It also shows the number of "silent" regions as marked by the annotators and the number of these regions that are preceded by IPBs. We see that Silence is by far the most significant correlate of prosodic phrasing in all corpora. However, the Italian dataset is very different from the others. It has the lowest rate of IPBs, the highest rate of IPB that precede pauses, and the lowest rate of silences with preceding IPBs. That is, it contains few IPBs, but almost all of these appear preceding a pause, although the rate of silence occurring with IPB is quite low. In the DIRNDL corpus, almost all IPBs (94%) occur at pauses. On the other hand, DUR, the Mandarin dataset, shows the opposite relationship; it has a relatively high rate of IPBs (37.10%), and from these only 61% precede pauses, the lowest rate in the table. The Italian corpus also shows a unique quality regarding intermediate phrase boundaries (indicated by a break index of 3) and silence (see Table 1). In this corpus we find that 95.66% of intermediate phrase boundaries occur at silences. This is in contrast to rates of 21.46%, 8.66% and 25.58% on BDC, DIRNDL and DUR material. Some intermediate phrase boundaries in the Italian data demonstrate qualities that would be strongly associated with intonational phrasing in English, including pre-boundary lengthening, silence, audible breath and pitch reset.

3. PHRASE BOUNDARY DETECTION

The detection task we focus on consists of deciding whether an intonational phrase boundary exists at the end of a word. We cast this as a binary classification task, for which we use AdaBoost [21] with stump hypothesis. All experiments were performed using the AuToBI toolkit for prosodic analysis [22]. The features used here are described in detail in [22] and in the IntonationalPhraseBoundaryDetectionFeatureSet distributed with version 1.2 of AuToBI at `http://speech.cs.qc.cuny.edu/autobi/`. They can be divided into four categories.

- **Pause** features: a boolean variable that indicates if the end of word precedes a silence and also duration of that pause.
- **Duration** features: the duration of the word and the difference of the duration of the current and following words.
- **Intensity** (dB) and **Pitch** (log Hz) contour features: these include the raw and speaker-normalized signals at different level of aggregations (mean, maximum, minimum and standard deviation). Speaker normalization is performed by z-score normalization.

We extract these values from each word, and then calculate their difference between the current and following word to create additional, context features.

4. CROSS-LANGUAGE PHRASE BOUNDARY DETECTION

The first set of experiments we present train a classifier using the training set from one language and test on the test set of another. We examine the F-score of the intonational phrase boundary class, since we find that overall accuracy provides little insight into differences in results. The results are shown in the left portion of Table 3. Every model achieves its highest F-score when tested on the DIRNDL corpus (except when the training and test corpus are the same). English and German produce particularly similar models, both reaching their best results on the other. This may be because they belong to the same language family. High performance on the DIRNDL corpus is probably due to the high correlation between IPBs and silence. The model that performs consistently worst results on every test corpus, and gets the worst results when used as a test corpus, is the Italian dataset. This may be explained by the anomalies discussed in Section 2.

In every corpus, silence is the most important feature for IPB prediction. While silence is an important predictor of phrasing, the use of this feature can overwhelm the role of other acoustic/prosodic features. To examine the power of cross-language IPB prediction without the use of silence, we repeat this experiment, omitting all

	Test Corpus							
	Full				Removed PPW			
Model	BDC	DIRNDL	DUR	Italian	BDC	DIRNDL	DUR	Italian
BDC	(0.79)	0.89	0.69	0.64	(0.00)	0.00	0.00	0.00
DIRNDL	0.79	(0.91)	0.71	0.65	0.00	(0,00)	0.00	0.00
DUR	0.74	0.80	(0.88)	0.64	0.18	0.45	(0.54)	0.40
Italian	0.43	0.61	0.61	(0.80)	0.00	0.04	0.01	(0.00)

Table 3. One vs One experiments F-Score results. Left columns show results for the full corpus and right columns show results after having removed pause-preceding words

pause-preceding words from the corpora. The results from these experiments are shown in the right portion of Table 3. When pause-preceding words are removed, we see that the predictive performance of almost all models drops to zero. One exception is in the DUR models, trained on the dataset with the lowest rate of IPBs preceded by a pause. This phenomena suggests that a) relevant information for IPB detection is extracted from the surrounding contexts of pauses and b) pause-related features have the most prediction power.

The second set of experiments uses a leave-one-corpus-out evaluation strategy, in which we train a classifier on three languages and test on the fourth, using the complete corpora in each case. Results are shown in Table 4. The results achieved on every test corpus are

	Baseline	Full		Remove PPW	
Test Corpus	Acc.	Acc.	F-Score	Acc.	F-Score
BDC	0.84	0.93	0.75	0.84	0.07
DIRNDL	0.88	0.98	0.92	0.89	0.19
DUR	0.63	0.83	0.73	0.63	0.00
Italian	0.89	0.90	0.67	0.90	0.33

Table 4. Leave-One-Out experiments

comparable or even better than the best results achieved on the one-versus-one experiments, ignoring the within-language results. When compared to the within-language results, DUR and Italian underperform significantly. We believe the poor performance on Italian is due to the unique relationship between pausing and intermediate phrase boundaries described in Section 2. We look again at the performance of our classifiers after removing pause-preceding words. Here we find that performance is only slightly better than the majority class baseline. However, we find that, when the DUR corpus is included in the training set, performance can exceed the baseline. This finding can be observed regardless of the evaluation language. This suggests that simply having more examples of phrasing without silence *regardless of language* can improve the prediction of this phenomenon. Presumably language differences would be more evident if the corpora included more instances of phrases that are not coincident with silence.

Because silence plays such a dominant role in detecting phrase boundaries, we repeat the leave-one-corpus-out experiments using specific feature subsets. The subsets are silence, pitch, intensity, duration and all-non-silence features. The results are shown in Table 5. We observe F-scores that are as good as those achieved by the all-feature classifier in bold. Except for the case of the model trained on English, Mandarin and German and tested on Italian, silence is the most relevant feature. In this experiment, the inclusion of silence features seems to worsen the accuracy of the classifier trained using the remaining features. We again attribute this to the relationship of silence and intermediate phrasing described in Section 2. When

Test Corpus	Silence	All\Silence	I	F0	Duration
BDC	**0.75**	0.55	0.52	0.19	0.02
DIRNDL	**0.93**	0.69	0.61	0.27	0.02
DUR	**0.73**	0.55	0.58	0.40	0.25
Italian	**0.67**	**0.70**	**0.69**	0.61	0.36

Table 5. Leave-One-Out results (F-Score) by feature subset.

we inspect the predictive power of these feature subsets, we find that intensity features yield the most predictive power, followed by pitch features, with duration features performing significantly worse. It is possible that the observed discriminative power of pitch is impacted by the use of Mandarin Chinese material in these experiments. The lexical tone present in the DUR corpus may lead to particularly different indications of phrasing when compared to the English, German and Italian material.

To test whether the *silence* classifiers are equivalent to the *all features* classifiers – that is, to see whether additional non-silence features add any extra predictive power – we use McNemar's test. We perform this test between the leave-one-out models trained using every feature and those trained using exclusively silence features. The test is applied to the contingency tables and determines whether the marginal probabilities are statistically the same. Table 6 shows that all the models except BDC-DIRNDL-Ita show significant differences, having distinct marginal probabilities (ie. classifications are different enough) with confidence at least 95%. This indicates that

Test Corpus	McNemar χ^2	p
BDC	5.6953	0.0170
DIRNDL	9.0313	0.0027
DUR	1.2308	0.2673
Italian	11.0769	0.0009

Table 6. McNemar's Chi-square and p values. Critical value at 95% significance level $\alpha = 0.05$ is 3.8415

the inclusion of the DUR dataset, which has the lowest rate of IPBs preceding pauses, in the training set leads to more complex phrasing classifiers.

5. WITHIN-LANGUAGE FEATURE ANALYSIS

In this section we look at the within-corpus prediction performance of each feature subset. The results of these experiments are shown in Table 7 and Figure 1. Table 7 shows F-score values for each feature subset. We see the relative reduction of error among the different classifiers compared with a majority baseline classifier. Figure

Dataset	All	Silence	All\Silence	I	F0	Duration
BDC	0.79	0.80	0.64	0.51	0.53	0.49
DIRNDL	0.91	0.91	0.62	0.55	0.20	0.00
DUR	0.88	0.71	0.84	0.81	0.40	0.39
Italian	0.80	0.77	0.69	0.71	0.48	0.40

Table 7. Within-language F-Scores values using feature subsets.

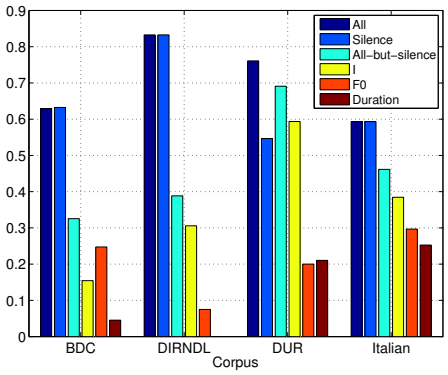

Fig. 1. Relative error reduction using feature subsets.

1 shows similar patterns of reduction of error from baseline among BDC, DIRNDL and Italian corpora: We see that using All and Silence features yields similar (and the maximum) reduction of error, followed by the All-but-silence feature set. Duration provides the least (if any, in the case of DIRNDL) reduction of error. However the DUR bar graph shows a different profile: it is the only corpus for which Silence fails to provide the highest error reduction while All-but-silence and I achieve the greatest reduction. Moreover, we find that only in BDC (SAE) are pitch features able to predict phrasing more reliably than intensity: on all other languages, intensity is a more reliable predictor. The importance of duration also varies by language: in DUR (Mandarin) and Italian, the predictive power is approximately equal to that of pitch, while there is almost no reduction of error based on duration in BDC and DIRNDL material.

6. COMPARING FEATURE DISTRIBUTIONS

In this section we examine the feature distributions of each language to gain more insight on the cross-language prediction performance. Specifically we look at the values of all the **Silence** and **Duration** features and the cross-word difference of speaker normalized mean values of **Pitch** and **Intensity**. Table 8 shows the mean and standard deviation values of each feature per class in each of the corpus. We find that there are some similarities in the differences in relevant features at phrase boundaries and non-phrase boundaries. The directionality of the differences are consistent in each languages. Phrase boundaries are more likely to precede silence and they precede longer silences. We observe increases in intensity and decreases in pitch across phrase boundaries compared to other word boundaries. We also observe that phrase ending words are longer than phrase-internal words. While the specific parameters of these relationships vary by language, the directionality is consistent. To compare the differences in the specific parameters, we calculate the average Kullback-Leibler divergence values for each pair of languages. These results, reported in Table 9, confirm the similarity of BDC and

Corpus	feature	IPB	not IPB
BDC	precedesSilence	0.55±0.84	-0.94±0.35
	follPause	2.86±17.02	0.01±0.05
	norm mean(I)	-0.36±0.74	0.06±0.76
	norm mean(f0)	0.13±0.75	-0.03±0.71
	duration	0.42±0.15	0.20±0.12
	dur(w_2)-dur(w_1)	-0.19±0.20	0.04±0.18
DIRNDL	precedesSilence	0.88±0.48	-0.95±0.32
	follPause	2.37±13.46	0.00±0.04
	norm mean(I)	-0.35±0.86	0.05±0.64
	norm mean(f0)	0.22±0.72	-0.03±0.58
	duration	0.53±0.21	0.37±0.24
	dur(w_2)-dur(w_1)	-0.24±0.28	0.03±0.35
DUR	precedesSilence	0.22±0.98	-0.90±0.43
	follPause	18.32±42.69	0.00±0.03
	norm mean(I)	-0.32±0.69	0.13±0.59
	norm mean(f0)	0.03±0.62	-0.07±0.56
	duration	0.55±0.23	0.50±0.23
	dur(w_2)-dur(w_1)	0.03±0.34	-0.01±0.28
Italian	precedesSilence	0.97±0.24	-0.74±0.67
	follPause	27.67±49.97	0.11±2.87
	norm mean(I)	-0.35±0.85	0.03±0.76
	norm mean(f0)	0.26±0.75	-0.04±0.64
	duration	0.71±0.25	0.36±0.25
	dur(w_2)-dur(w_1)	-0.34±0.36	0.04±0.38

Table 8. Mean and std. dev. of example features from the four feature sets

DIRNDL. The Italian corpus proves to be the most dissimilar with

Corpus	BDC	DIRNDL	DUR	Italian
BDC	0.00	0.13	0.36	0.62
DIRNDL	-	0.00	0.20	0.59
DUR	-	-	0.00	0.42
Italian	-	-	-	0.00

Table 9. Mean KL-divergence values for each pair of corpus.

respect phrasing boundaries to any other corpus. This is a concise description of the contrasts that we observe in Table 9 and Figure 1.

7. CONCLUSION AND FUTURE WORK

We have presented a simple approach to cross-language phrase boundary detection using one-on-one and leave-one-out experiments. Our experiments show that classifiers trained on several languages perform better over the same test corpora. We find that, while pause (silence) is the most important feature for IPB prediction in all languages examined, the relative importance of other features does vary significantly by language. We also find that different acoustic correlates of prosodic boundaries exist in different languages. In SAE, German and Mandarin, the relative importance of features is silence $>$ intensity $>$ pitch $>$ duration, while for English pitch is more important than intensity. These differences do not appear to be attributable to language family, since, e.g. English and German display different patterns.

In future work, we will examine more explicit methods of cross-language adaptation. We also will explore additional languages from different language families.

8. REFERENCES

[1] Zhongqiang Huang and Mary Harper, "Appropriately handled prosodic breaks help pcfg parsing," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, USA, 2010, HLT '10, pp. 37–45, Association for Computational Linguistics.

[2] Vladimir Eidelman, Zhongqiang Huang, and Mary Harper, "Lessons learned in part-of-speech tagging of conversational speech," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA, 2010, EMNLP '10, pp. 821–831, Association for Computational Linguistics.

[3] K. Chen, M. Hasegawa-Johnson, A. Cohen, S. Borys, S. Kim, J. Cole, and J. Choi, "Prosody dependent speech recognition on radio news," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 232–245, 2006.

[4] Yi Su and Frederick Jelinek, "Exploiting prosodic breaks in language modeling with random forests," in *Speech Prosody*, 2008.

[5] S. Maskey and J. Hirschberg, "Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization," in *Eurospeech*, 2005.

[6] Mohammed Hoque, Mohammad Sorower, Mohammed Yeasin, and Max Louwerse, "What speech tells us about discourse: The role of prosodic and discourse features in dialogue act classification," in *IJCNN*, 2007.

[7] Colin Wightman, Nanette Veilleux, and Mari Ostendorf, "Use of prosody in syntactic disambiguation: An analysis-by-synthesis approach," in *HLT*, 1991, pp. 384–189.

[8] Anton Batliner, Elmar Nöth, Jan Buckow, Richard Huber, Volker Warnke, and Heinrich Niemann, "Duration features in prosodic classification: why normalization comes second, and what they really encode," in *Prosody 2001*, 2001.

[9] Ken Chen, Mark Hasegawa-Johnson, and Aaron Cohen, "An automatic prosody labeling system using ann-based syntactic-prosodic model and gmm-based acoustic-prosodic model," in *ICASSP*, 2004.

[10] Kyuchul Yoon, "Teaching english intonation through learner utterances with cloned native intonation," *Modern Studies in English Language & Literature*, vol. 54, no. 1, pp. 147–171, 2010.

[11] Julia Hirschberg and Christine Nakatani, "Using machine learning to identify intonational segments," Tech. Rep. SS-98-01, AAAI, 1998.

[12] Vivek Rangarajan Sridhar, Srivinas Bangalore, and Shrikanth Narayanan, "Exploiting acoustic and syntactic features for prosody labeling in a maximum entropy framework," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 16, no. 4, pp. 797–811, 2008.

[13] Nanette Marie Veilleux, *Computational models of the prosody/syntax mapping for spoken language systems*, Ph.D. thesis, Boston University, Boston, MA, USA, 1994, Major Professor-Ostendorf,, Mari.

[14] Phillip Koehn, Steven Abney, Julia Hirschberg, and Michael Collins, "Improving intonational phrasing with syntactic information," in *ICASSP*, 2000.

[15] Chong-Jia Ni and Ai-Ying Zhang, "The comparison between mandarin break detection and english break detection," in *Pattern Recognition*, vol. 321 of *Communications in Computer and Information Science*, pp. 597–605. Springer Berlin Heidelberg, 2012.

[16] Andrew Rosenberg, Erica Cooper, Rivka Levitan, and Julia Hirschberg, "Cross-language prominence detection," in *Speech Prosody*, 2012.

[17] J. Hirschberg and C. Nakatani, "A prosodic analysis of discourse segments in direction-giving monologues," in *Proc. of the 34th conference on Association for Computational Linguistics*, 1996, pp. 286–293.

[18] C. Shih and B. Ao, "Duration study for the Bell Laboratories Mandarin text-to-speech system," in *Progress in Speech Synthesis*, van J. Santen, R. Sproat, J. Olive, and J. Hirschberg, Eds., pp. 383–399. Springer-Verlag, New York, 1997.

[19] Kerstin Eckart, Arndt Riester, and Katrin Schweitzer, "A discourse information radio news database for linguistic analysis," in *Linked Data in Linguistics*, 2012.

[20] "GToBI," `http://www.gtobi.uni-koeln.de/`.

[21] Yoav Freund and Robert E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and Systems Sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[22] Andrew Rosenberg, "Autobi - a tool for automatic tobi annotation," in *Interspeech*, 2010, pp. 146–149.