Web Harvest of Minimal Intonational Pairs

Jonathan Howell
Dept. of Linguistics
Cornell University
jah238@cornell.edu

Mats Rooth
Dept. of Linguistics and FCI
Cornell University
mr249@cornell.edu

Abstract

This paper describes experiments on gathering spoken-language data on the web that bears on issues of the phonetics-phonology and semantics-pragmatics of intonation. The target data are tokens of fixed word strings like "than I did", where intonation varies in a way which correlates with grammatical and pragmatic context. In a web harvest procedure, audio files were identified using a search engine based in speech-to-text, downloaded, and cut to a relevant segment under program control. In an application of such a database, an SVM classifier was trained to make a grammatically determined distinction in intonation based on purely acoustic cues. Sources of error in the retrieval are quantified.

1 Introduction

We are interested in collecting from web sources audio recordings of utterances that bear on theories of intonation. In particular, we would like to create databases of multiple repetitions of tokens embedding a fixed word string $w_1 \ldots w_n$, within which intonation varies in a way that correlates with syntax, semantics, and/or pragmatics. For instance, in comparative sentences such as (1a,b,c), there is an intuition that intonational focus in *than*-clause co-varies with the main clause in a systematic way. A generalization which turns out to be very robust (see Section 4) is that when reference varies in the subject position between the main and *than*-clauses as in (1a), the subject pronoun *I* in the *than*-clause is intonationally focused in the sense of Jackendoff (1972). When reference is constant in the subject position as in (1b) and (1c), the subject in the *than*-clause is unaccented.

1) a. She did more than I did.
   b. I wish I had done more than I did.
   c. I did more than I did last time.

The target sequence w_1,w_2,w_3 in this case is "than I did". In sentences (1a-c), this substring is constant, but intonation varies in a way that correlates with the grammatical context. (1a,b) is a minimal pair, where arguably a single parameter distinguishes the clauses [than I did] in the two utterances. As articulated in theories of the semantics of focus intonation such as Rooth (1991) and Schwarzschild (1999), and accounts of the phonology-phonetics of focus intonation such as Truckenbrodt (1995) and Féry and Samek-Lodovici (2006), this is a parameter which has both a semantic/pragmatic and phonological/phonetic interpretation.

Constructing indexed web corpora in which such pairs could be retrieved, or collecting large samples of given minimal pairs from web sources, could allow both the semantic/pragmatic conditioning of the intonation and its phonetic realization to be studied and modeled on an unprecedented scale. Linguistic theories of intonation ultimately capture correlations between acoustic form and syntax, semantics and pragmatics; they make predictions about what prosodic patterns fit into what grammatical and pragmatic contexts. We would like to confront deep, logically formalized theories of this correlation with massive amounts of data harvested on the web.

This paper describes experiments in which samples for several targets were collected using a web harvest. Section 2 explains the harvest method. Section 3 evaluates the efficacy of the retrieval, discussing sources of error such as failure to retrieve an audio file over the network, and speech recognition errors. Section 4 describes an application of the data sample, where an SVM classifier was trained to make a semantically motivated distinction in the location of contrastive focus based on acoustic parameters. Section 5 gives information about additional samples being collected, and the final section offers our conclusions and suggestions about the form of web corpora of spoken language data that would be suitable for research on intonation.

2 Web harvest method

We used an external search engine with indexing based on automatic speech recognition to identify of the URLs of audio files that contain (or may contain) tokens of the target word sequence $w_1 \ldots w_n$. We aimed to use a basic approach of downloading html pages from the search engine, using simple text processing to extract URLs of audio files and other relevant information, retrieving and cutting audio files with software with a command-line interface, and using makefiles and glue languages to control the retrieval and integrate the software components.

Kohler *et al*. (2008), which discusses technology and applications for retrieval of spontaneous conversational speech, lists online search engines that index spoken language. Our survey indicated that Everyzing (search.everyzing.com) is suitable for our experiment in the following respects:

i. Searches for word strings are possible in the query language, including strings involving frequent words (stop words).
ii. Initial experimentation indicated that enough data is indexed to retrieve hundreds or thousands of tokens of the strings we are interested in.
iii. The indexed material includes a large amount of conversational data, where intonational phenomena of interest are common, and utterances are produced naturalistically.
iv. In addition to the URL of an audio file, the search engine returns time offsets for each target word. This makes it possible to automate cutting the audio files.
v. Initial experimentation indicated that, for target strings of interest, the accuracy of the engine's speech recognition was good.

Everyzing indexes both pure audio files and files with combined video and audio. Since the size of the files to be retrieved was an issue, we restricted the experiment to audio files to minimize file size. These audio files are always in mp3 format.

An experimenter first queried the engine in a browser, in order to determine whether a given string is common enough. After this, the retrieval is performed under program control, in a sequence that mimics what a human would do in interacting with the engine through a web browser.

For retrieving material from the search engine, we used curl 7.16.3, which is a command line

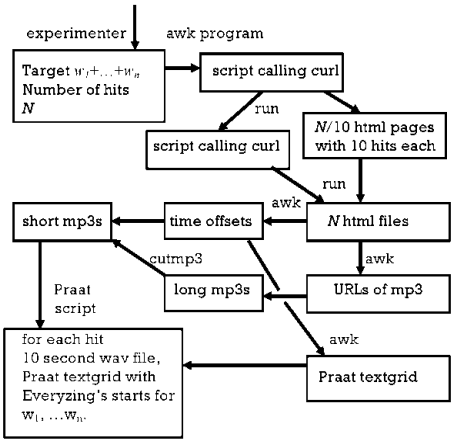

Figure 1. Workflow for mp3 retrieval and editing.

tool that retrieves data designated in URL syntax (Stenberg, 2008). The inputs to the procedure, which is diagrammed in Figure 1, are the target string and the number *N* of hits to be retrieved.

The first programmatic step constructs a shell program which contains *N*/10 calls to curl. Each involves a URL that embeds the target word string in the format "w_1+…+w_n" and an integer which functions as an index into the sequence of hits. Such a string is equivalent to the URL of the page that Everyzing displays when asked in the browser to display a group of 10 hits. Running the shell scripts retrieves *N*/10 html files, each representing 10 hits, and writes another shell script used in the next step. That script calls curl *N* times, retrieving html files for individual hits. At this point, processing with awk extracts from each file the URL of an mp3, and time offsets for the individual target words in the audio file.

Audio files are retrieved with curl, and subsequenty cutmp3, a command line program for cutting mp3 files, is used to cut a 10-second audio file from each long mp3 file, referring to the time offset (Puchalla, 2008).

Finally, we prepared data for analysis in the phonetic software package Praat (Boersma and Weenink, 2001). Mp3 files were converted to wav format, and using the time offsets of the target words, a Praat TextGrid file was prepared, which aligns the acoustic signal with the target words. Bit rate in the "than I did" dataset varied from 32 to 256 kbits/s and sampling frequency 11025 to 44100 Hz. By comparison, speech files in the often used Switchboard corpus were recorded over the telephone at 8 kbits/s and with a sample rate of 8000 Hz. Note that mp3 is a lossy

inmyopinion350.hits	html for hits 350-359
inmyopinion360.hits	html for hits 360-369
inmyopinion351.hit	html for hit 351
inmyopinion352.hit	html for hit 352
inmyopinion352.mp3name	URL of audio file
inmyopinion352.cut	time offset for hit 352
inmyopinion352.mp3	long audio file of hit 352
inmyopinion352-b.mp3	10-second audio file of hit 352

Table 1. Files from a retrieval with target "in my opinion".

compression format, which could have an impact on subsequent processing of the audio signal; however these are the available data.

In the scripts that issue requests to search.everyzing.com, we used a time delay of 25 seconds between the termination of one curl retrieval and the issuance of the next, to avoid flooding the server. We found that the audio files retrieved from various sources were often very long, and that retrieval of audio files would sometimes hang; therefore we imposed a time limit of 600 seconds for retrieving each audio file.

Files created in a retrieval run for "in my opinion" are exemplified in Table 1. The file inmyopinion352.mp3 is the full audio signal, while in inmyopinion352-b.mp3 signal has been cut to a 10-second interval flanking a putative occurrence of the target.

In the in-my-opinion run the long mp3 files had a median size of 20MB, and a maximal size of 180MB for a two hour and five minute recording of a university forum. The total size of 714 mp3s retrieved in this run is 16.4GB. The run took 24 hours.

Table 2 lists the most common domain names, indicating a predominance of radio content. WEEI, WNYC, KPBS, and WRKO are radio stations; White Rose Society is an archive of progressive radio; the items in the akamai domain comprise three AM radio stations; NPR is National Public Radio. Podtrac is site that matches podcast and advertising content.

116	a1135.g.akamai.net
110	hosted-media.podzinger.com
76	media.weei.podzinger.com
58	feeds.wnyc.org
54	media.libsyn.com
51	podcastdownload.npr.org
50	feeds.feedburner.com
39	library.kraftsportsgroup.com
33	www.whiterosesociety.org
24	www.kpbs.org
21	www.podtrac.com
21	media.wrko.podzinger.com

Table 2. The most frequent domain names in the in-my-opinion run.

3 Evaluation of retrieval efficacy

In a pilot experiment conducted prior to full implementation of the procedure described in Section 2, 179 purported tokens of the string "than I did" were downloaded manually by the experimenter via Everyzing and cut manually using Praat. 91 were identified as unique true occurrences of the target.

In one of several subsequent harvests using the procedure described in Section 2, 2,300 tokens of the target string "he himself" were reported by the search engine, and *N* was set at 300. The shell scripts retrieved 30 html files representing 300 hits, and then retrieved 285 individual hit html files. From these, awk generated 263 files with time-offset information (22 contained no time-offset information). 60 of the 285 mp3 files downloaded were unreadable. Upon further investigation, many of the unreadable files were in fact recoverable by a new search of Everyzing with uniquely identifying text and then manual download. This suggests corruption during the curl retrieval, rather than a corrupt file at the source.

An experimenter listened to all short mp3 files individually and those not containing unique occurrences of the target utterance were rejected. In 16 cases, the cut file contained inaccurate time-offsets, resulting in a short mp3 file that did not contain the purported target. Often this was due to sponsorship information in public radio podcasts which was appended to the mp3 file but did not appear in the Everyzing media player or transcription. In 25 cases, a rejected file contained an incorrectly transcribed token with a near match (e.g. *sees himself, um himself*, *eek himself*, *has himself*) or sometimes with nothing resembling the target (e.g. *building stuff*, *purify*, *independent senator*). Four of the short mp3 files were duplicates of previous files. The remaining true, unique tokens of the target numbered 154, roughly one half of the set initially queried. Other retrieval runs yielded comparable, although different results, as summarized in Figure 2.

We close this section with a comparison of the size of the datasets that can be harvested on the web with a hand-annotated speech corpus. Switchboard (Godfrey et al., 1992) contains 240 hours of speech from 2400 telephone conversations, a third of which has been made available

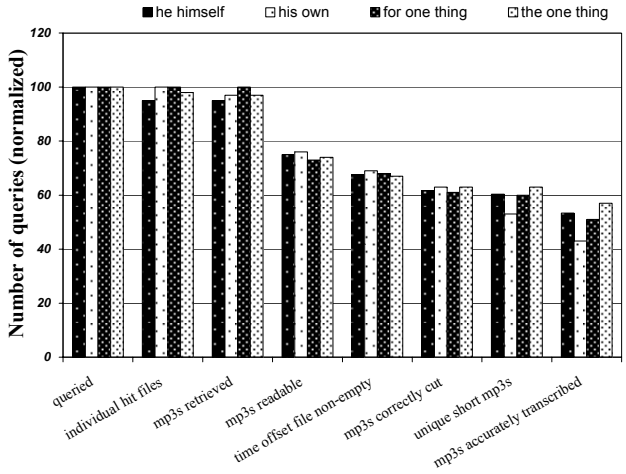

Figure 2. Detailed retrieval efficacy at different processing stages compared for 4 different retrieval runs: (normalized to 100, n=300, 100, 100, 100).

by Calhoun et al. (2005) with annotation for syntactic structure as part of the Penn Treebank (Marcus et al., 1993), dialog acts (Shriberg et al. 1998) and information status (Calhoun et al., 2005) and has formed the basis of numerous studies relating prosody, syntax and semantics (cf. Bell et al., 2009; Calhoun, 2006, 2007, 2008; Sridhar et al., 2008, Nenkova and Jurafsky, 2007; Jurafsky et al., 1998). Clearly, this type of static, richly annotated corpus offers many virtues, particularly as a standard of comparison.

Unfortunately, the restricted size of such a corpus due to the limitations of human resources means that it is not always large enough to allow statistical analysis of specific linguistic constructions. The Switchboard-1 corpus available at the Linguistic Data Consortium Online contains 26,151,602 word tokens. Figure 3 compares, for each of five targets, (a) the number of tokens contained in the Switchboard sample (b) the number of true tokens we have already collected and verified from Everyzing, and (c) the projected number of true tokens from Everyzing based on the number of hits returned and assuming a roughly 50% retrieval efficacy. While the Switchboard data may prove a useful baseline for certain target expressions, it is clear that a dynamic web harvested corpus will be not only less costly but much greater in scope. In particular, this allows us to apply machine learning techniques as an alternative to prosodic annotation by human experimenters which necessarily introduces certain theoretical assumptions such as the prosodic ontology of the Tones and Breaks Indices (TOBI) framework (Silverman et al.,1992) for prosodic annotation.

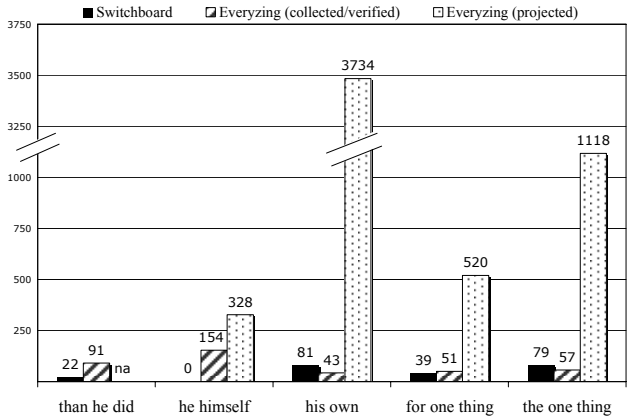

Figure 3. Comparison for each target expression of (a) number of tokens in the Switchboard corpus, (b) number of good tokens already collected and identified in the web-harvested corpus and (c) the number of projected tokens available through Everyzing at the time of harvest, based on total hit count and assuming 50% retrieval efficacy.

4 Machine learning classification

This section describes an experiment which illustrates the scientific interest of the web samples, and shows that it possible to obtain consistent results with these samples, despite variation in discourse type, recording conditions, and signal parameters, and despite the possibility of the lossy mp3 format interfering with audio processing.

On many semantic theories, unaccented material must be licensed anaphorically. In practice, however, such linguistic antecedents are not always available in the discourse; they may be inferable from the non-linguistic context.

While corpus data have the virtue of naturalness, they show extreme variation with respect to discourse context. (Laboratory-elicited data, by contrast, may be artificially controlled for discourse context although in that case the design is necessarily constrained by the experimenters' theoretical assumptions). The comparative construction discussed in Section 1 is subject to this variation, yet it has the virtue of encoding, for any given instance, an explicit antecedent. The scope of the focus (focus indicated with subscript F) is the *than*-clause, and the antecedent is contained with the main clause.

2) a. He stayed longer than $[I]_F$ did.
   antecedent: *He stayed x long*
   b. I should have liked that song a lot more than I $[did]_F$.
   antecedent: *I should have liked that song x much*
   c. I understand even less than I did $[before]_F$
   antecedent: *I understand even x little*

When the subject of the antecedent matrix clause varies from the subject of the embedded clause, theory predicts intonational prominence on *I*. When the subjects corefer, theory predicts reduced prominence. In the experiment, we trained a classifier to discriminate these two categories given only acoustic information.

As described in Section 3, we collected 179 purported tokens of the string "than I did". Each of the short sound files produced was then annotated into segments using Praat: the vowels of *than*, *I* and *did*, as well as the stop duration in *did*. Praat scripts were then used to extract 308 acoustic parameters (see Howell and Rooth, 2009).

Each token and its preceding environment was transcribed by hand. From this text, the tokens were manually classified by an experimenter into the two semantico-grammatical categories. When the subject of the main clause and the *than*-clause (i.e. *I*) varied, tokens were categorized into a class *s* (subject focus: 46/91 tokens). When the subject of the main and *than*-clauses remained constant and some contrastive post-verbal material (e.g. a temporal phrase) followed (36/91 tokens) or when the subject of the main clause and *than*-clauses remained constant and no contrastive material followed (focus on *did*: 9/91), tokens were categorized into a class *ns* (non-subject focus: 45/91). This classification can be made by grammatical and semantic criteria, and is nearly uncontroversial.

Model 1: 82.4%

	predicted	
true	s	ns
s	35	5
ns	11	40

Model 2: 79.1%

	predicted	
true	s	ns
s	34	7
ns	12	38

Model 3: 89.0%

	predicted	
true	s	ns
s	44	8
ns	2	37

Model 4: 92.3%

	predicted	
true	s	ns
s	43	4
ns	3	41

Table 3. Contingency tables and total accuracies for predictions of different SVM classifiers using OHOCV for binary classification of subject and non-subject conditions.

A supervised support vector machine (SVM) classifier was trained in the R statistical computing environment (R Development Core Team, 2008) using an installation of the libsvm library (Chang and Lin, 2001) in package e1071 (Dimitriadou et al., 2009), using the two classes *s* and *ns*. The classifier was run with all 308 acoustic parameters (Model 1) on the 91 tokens categorized as *s* and *ns*. The success of the classifier is measured according to a one held out cross-validation (OHOCV) test. One of the 91 tokens is held out and the classifier is trained on the remaining 90. This is repeated for all of the tokens and a total accuracy is calculated on the number of successful classifications. Model 1 achieved a total accuracy of 82.4% (16 misclassifications). The results for this and following models are summarized in Table 3. A second classifier (Model 2) was tried with only 212 parameters, those extracted from *I* and *did* only, which performed marginally worse at 79.1% (19 misclassifications).

Next, we attempted different feature selection methods including a backwards-elimination

technique using a random forest classifier in the R package varSelRF (Diaz-Uriarte, 2009). This produced an optimal decision tree with just a single variable: the duration of *I*. An svm classifier with just this variable (Model 3) achieved a total accuracy of 89.0% (10 misclassifications). Finally, we added to this variable the closure duration for the onset of *did*, and the difference in first and second formants at 40% of I (4*(total duration)/10) yielding a best model (Model 4) with 92.3% total accuracy (7 misclassifications).

These results offer strong empirical support of the theoretical prediction: coreference of the subject is highly correlated with reduced acoustic prominence and lack of coreference is highly correlated with increased acoustic prominence. Morover, a small set of cues for the categories involving duration and vowel quality, and not involving pitch, is sufficient to distinguish the categories acoustically.

It is not obvious that the correlation between acoustic form and semantic-grammatical context should hold up so well in such a diverse sample. We anticipate that some correlations discussed in the literature will be disconfirmed when tested against large samples harvested on the web, while others (like this one) will be confirmed and quantified.

5 Additional targets

Several other data harvests are planned or in progress. Since the machine learning classification in Section 4 revealed segmental information, in particular formant extrema, to be relevant in the detection of focus placement, we plan to harvest other targets within the same comparative paradigm, yet with different vowels: *than he did* [ij], *then they did* [ej], *than you did* [uw], *than it did* [ ]. Featural enhancement models predict that segmental features should also inform the focus placement classification for tokens with these vowels. If this is correct, one could build a successful classifier by providing information about vowel identity.

The retrieval of targets *he himself* and *his own* mentioned in Section 3 forms part of a larger harvest of targets, including other intensive reflexives, alleged to have an invariant focus pattern (e.g. Cantrall 1973; Creswell 2002; König and Gast 2006). One possible approach follows the semi-supervised method used for the comparative targets, with potentially controversial human classification into different intonational categories (e.g. *HE HIMSELF, he HIMSELF*). Another approach is to apply unsupervised machine learning to identify different classifications independent of human perception.

Accent type will be investigated using minimal pairs where syntax favors a particular accent. For example, most occurrences of the target *for one thing* have a "topic" accent (L*H in TOBI annotation) while most occurrences of the target *the one thing* have a "focus" accent (H*), the two predicted to differ in pitch contour. Other configurations occur with accent placement on other constituents (e.g. *except for one THING*, *that's the one THING*). The intension is to train a classifier on these less controversial targets and then to apply it more widely to occurrences of *one thing* generally.

These targets illustrate the value of working with a very large source of data. It is possible to obtain non-trivial datasets for phenomena which, though they do not strike speakers of English as exotic, are in fact rare.

6 Discussion

We have established by example that large samples of spoken-language phenomena can be gathered on the web using simple web retrieval, text processing, and audio processing methods. The procedure is cheap. Attempted retrieval of 1000 potential tokens results in retrieval of about 750 audio files, containing hundreds of actual tokens of the target. A run of this size requires network transfer and storage of about 20GB of data. Disk capacity for this volume of data costs a few dollars. Network charge environments are readily available where transfer costs for this volume of data is on the same scale. Since the retrieval is done under program control, cost in experimenter time is also small.

The analysis in Section 3 shows that the quality of the retrieved samples varies with the target. Thinking of the system as a prototype concordance interface that presents a list of 10-second audio segments to the linguist for examination, a proportion of 50% of segments that actually contain the target seems acceptable.

It is natural to wonder whether any of the hand work in the SVM classification procedure can be automated. These steps are:

(i) Transcription of the 10-second segment.
(ii) Temporal word alignment in Praat.
(iii) Alignment of sub-phonemic acoustic events in Praat.
(iv) Classification into the semantic-grammatical categories *s* and *ns*.

Automation of any of the steps would speed up creating a dataset. Given a word transcription, there are available solutions for creating a word level alignment. For instance Yuan and Liberman (2008a,b) used a forced aligner based on the HTK HMM toolkit to create a Praat text grid with work alignments, given a word transcription. It seems likely that the same technique would be usable in (iii). This would allow the acoustic-phonetic hand work to be automated, with the additional advantage of making that work replicable.

Search.everyzing.com went offline in June 2009. Various large sites with indexing bases on speech recognition are online, such as Fox Business News and WNYC. While Google's Gaudi offering is still limited to material from the US presidential election, this could in the future be a replacement generic audio search offering.

An interesting angle is provided by individual sites that intend to expose their multimedia material to generic text search by providing transcriptions. For instance audio.weei.com (an Everyzing customer) has pages containing en embedded player for sports radio programs with functionality for search within a radio program, an mp3 download option, and a transcription. Given a list of sites, the tokens can be found with a generic text search engine, or with a textual search engine API.

The current reality is that creating datasets of sufficient size requires interacting with numerous different sites, each with its own HTML representation. Thus the text-processing work that extracts the URL of the mp3 and a time offset would have to be implemented many times, once per site. This could be compensated for by using a more sophisticated scraping technology which works with the Document Object Model representation of the page, rather than simply the string representation like the procedure in Section 3. We hope to look at available systematic solutions to this problem.

A bottleneck in the current procedure is the need for an experimenter to listen to the hits in order to select the actual tokens and create a corrected transcription of the host sentence. This is not really onerous if one is working with a few hundred examples, and at some point we want to evaluate the data as linguists anyway. But suppose 10,000 candidate tokens were available; having to listen to about 5000 incorrect tokens just to reject them would be a waste of time. We plan to look at building a targeted classifier that, for a single target, attempts to sort out the correct candidates from the incorrect ones. The classifier would be bootstrapped from a manually classified subset. This classification problem is similar to keyword spotting (e.g. Keshet et. al. 2009).

On top of general objections to basing linguistic research on commercial search engines (Kilgariff 2007), in our procedure there are sources of bias in the automatic speech recognition. It seems plausible that a speech recognizer could have substantially different recall rates for two phrase types with the same word string, but different prosodic patterns. If so, the samples collected would be biased in a way that could easily affect the evaluation of linguistic hypotheses. While it is not possible to avoid this problem within our architecture, one should try to quantify it. This might be done by finding recordings where a correct transcription is independently available. Or if working with a generic search engine, one could put test data onto the web, and measure the recall of the engine for the specific prosodic realizations of the target.

Our results and experience are suggestive about suitable forms of indexing for a web corpus of spoken language. As described in Section 3, searches for fixed word strings are useful in finding data bearing on issues on the realization and conditioning of intonation. Such searches appear to compensate for deficiencies in speech-to-text technology, because accuracy at the scale of a short tuple can be good, even if coherent transcriptions are not produced at the sentence scale. Thus it seems attractive to create web corpora of spoken language indexed by word n-grams, combined with a query system including variables and disjunctions. This would parallel web corpora and concordancing tools for written data (Fletcher, 2007).

Our results also suggest the feasibility of automatically indexing spoken-language corpora by prosodic features. Assuming that the classification results from Section 3 extend to general contexts, an SVM classifier is able to classify tokens of the first person pronoun "I" as focused or not as well as a human, based on local, paradigmatic signal features. This could make it possible to index a corpus automatically with a limited number of prosodic features.

References

Alan Bell, Jason Brenier, Michelle Gregory, Cynthia Girand, and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60(1):92-111.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5:341–345.

Sasha Calhoun. 2006. *Information Structure and the Prosodic Structure of English: a Probabilistic Relationship*. PhD thesis, University of Edinburgh.

Sasha Calhoun. 2007. Predicting focus through prominence structure. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.

Sasha Calhoun. 2008. Why do we accent words? The processing of focus and prosodic Structure. Presented at *Experimental and Theoretical Advances in Prosody*, Cornell University, NY.

Sasha Calhoun, Malvina Nissim, Mark Steedman and Jason Brenier. 2005. A framework for annotating information structure in discourse. In *Frontiers in Corpus Annotation II: Pie in the Sky*. ACL2005 Conference Workshop, Ann Arbor, MI.

William R. Cantrall. 1973. Why I would relate 'own', emphatic reflexives, and intensive pronouns, my own self.' *Papers from the Ninth Regional Meeting*, eds. C. Corum, T.C. Smith-Stark and A. Weiser, 57-67. Chicago: Linguistic Society.

Chih Chang, and Chih Lin. 2001. LIBSVM: a library for support vector machines. URL http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Cassandre Creswell. 2002. The use of emphatic reflexives with NPs in English. In *Information Sharing*, eds. K. van Deemter and R. Kibble. Stanford, CA: CSLI Publications.

Ramon Diaz-Uriarte. 2009. VarSelRF: variable selection using random forests. URL http://ligarto.org/rdiaz/Software/Software.html, R package version 0.7-1.

Evgenia Dimitriadou, Kurt Hornik, Friedrich Leisch, David Meyer and Andreas Weingessel. 2009. *e1071: Misc functions of the department of statistics (e1071)*, TU Wien . R package version 1.5-19.

Caroline Féry and Vieri Samek-Lodovici. 2006. Focus projection and prosodic prominence in nested foci. *Language* 82:131-150.

William Fletcher. 2007. Implementing a BNC-Comparable Web Corpus. *Web as Corpus* 3.

John J. Godfrey, Edward Holliman and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *IEEE ICASSP-92. ACL Workshop on Discourse Annotation*.

Jonathan Howell and Mats Rooth. 2009. A corpus search methodology for focus realization. Poster presented at the 157th Meeting of the Acoustical Society of America, Portland, OR. http://hdl.handle.net/1813/13093.

Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.

Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William Raymond. 1998. Reduction of English function words in Switchboard. *Proceedings of ICSLP-98* 7.

Joseph Keshet, David Grangier, and Samy Bengio. 2009. Discriminative Keyword Spotting. *Speech Communication* 51(4):317-329.

Adam Kilgariff. 2007. Googleology is bad science. *Computational Linguistics* 33(1):147-151.

Joachim Kohler, Martha Larson, Franciska de Jong, and Wesse Kraaij. 2008. Spoken content retrieval: searching spontaneous conversational speech. *SSCS* 2008.

Ekkehard König and Volker Gast. 2006. Focused assertion of identity: A typology of intensifiers. *Linguistic Typology* 10.

Mitchell P. Marcus, Beatrice Santorini and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: the Penn treebank. *Computational Linguistics* 19:313-330.

Ani Nenkova and Dan Jurafsky. 2007. Automatic detection of contrastive elements in spontaneous Speech. In *Proceedings of the IEEE workshop on Automatic Speech Recognition and Understanding* (ASRU), Kyoto, Japan.

Jochen Puchalla. 2008. Cutmp3. URL http://www.puchalla-online.de/cutmp3.html.

R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Mats Rooth. 1991. A Theory of Focus Interpretation. *Natural Language Semantics*, 1(1).

Roger Schwarzschild. 1999. Givenness, avoid F and other constraints on the placement of focus. *Natural Language Semantics*. 7(2):141-177.

Elizabeth Shriberg, Rebecca Bates, Andreas Stolcke, Paul Taylor, Daniel Jurafsky, Klaus Ries, Noah Coccaro, Rachel Martin, Marie Meteer, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech* 41: 443-492.

Kim Silverman, Mary Beckman, John Pitrelli, Mari Ostendorf, Colin W. Wightman, Patti Price, Janet Pierrehumbert & Julia Hirschberg. 1992. A standard for labelling English prosody. *ICSLP*.

Vivek Kumar Rangarajan Sridhar, Ani Nenkova, Shrikanth Narayanan and Dan Jurafsky. 2008. Detecting prominence in conversational speech: pitch accent, givenness and focus.In *Proceedings of Speech Prosody*, Campinas, Brazil.

Daniel Stenberg. 2008. *cURL and libcurl*, http://curl.haxx.se.

Hubert Truckenbrodt. 1995. *Phonological Phrases-their Relation to Syntax, Focus, and Prominence*. PhD thesis, MIT.

Jiahong Yuan and Mark Liberman. 2008a. Speaker identification in the SCOUTUS corpus. *Journal of the Acoustical Society of America*.

Jiahong Yuan and Mark Liberman. 2008b. Vowel acoustic space in continuous speech: an example of using audio books for research. *CatCod 2008*.