

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 $w_1 \dots w_n$, 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 covaries 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 Jackendoff (1977). 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. He did more than he did before.

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 conclu-

sions 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 speech-to-text technology to identify of the URLs of audio files that contain (or may contain) tokens of the target word sequence $w_1 \dots w_n$. 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 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.

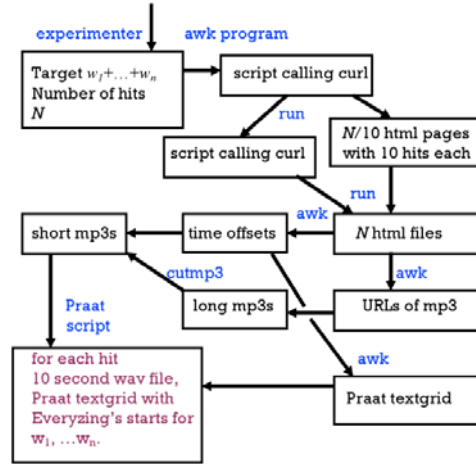


Figure 1. Workflow for mp3 retrieval and editing.

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 “ $w_1 + \dots + w_n$ ” 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 subsequently 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.

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.

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”.

In the in-my-opinion run the long mp3 files had a median size of 20MB, and a maximal size of 180 MB 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. WEEL, 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.

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

116	al135.g.akamai.net
110	hosted-media.podzinger.com
76	media.weel.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.

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, 300 tokens of the target string “he himself” were queried. 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 which had been automatically generated numbered 154, roughly one half of the initial queried.

Other retrieval runs yielded comparable, although different results, as summarized in Figure 2.

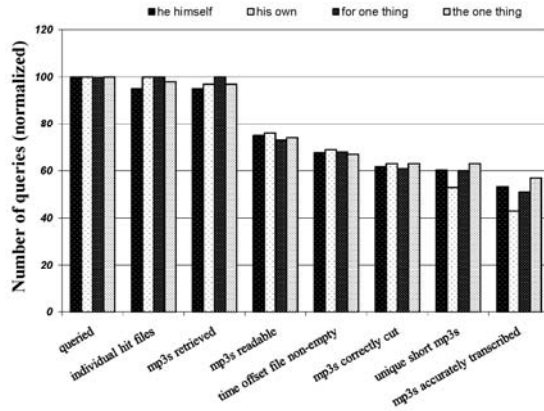


Figure 2. Detailed retrieval efficacy at different processing stages compared for 4 different retrieval runs: (normalized to 100, n=300, 100, 100, 100).

4 Machine learning classification

This section describes an experiment which illustrates the scientific interest of the web samples, and indicates the feasibility of prosodic indexing of web corpora.

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 (compared to laboratory-elicited data, for example, which can be controlled for discourse context yet also offer quantifiability). The comparative construction discussed in Section 1 is subject to this variation, yet it has the feature 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 less*

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 this 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, including values for duration, intensity, energy, amplitude, f0, and vowel formants. Formant bandwidths and the first three harmonics were also collected as measures of spectral tilt or balance (see Sluijter and van Heuven, 1996 and Campbell and Beckman, 1997).

Mean, extrema and range were collected for most continuous measures. The time at which an extrema occurred relative to the vowel duration was taken, and in some cases, a measure was also taken at the time corresponding to the maximum of another measure (e.g. the value of f1 was taken at the time corresponding to the f0 maximum). Each token and its preceding environment was transcribed into prose 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, but some post-verbal material (e.g. a temporal phrase) contrasted (following focus: 36/91 tokens) or when the subject of the main clause and *than*-clauses remained constant and no material followed (focus on *did*: 9/91), tokens were categorized into a class “ns” (non-subject focus: 45/91).

From this, a supervised support vector machine (SVM) classifier was trained in the R statistical computing environment (R Development Core Team, 2008). Introduced by Cortes and Vapnik (1995) for binary classification, SVM is increasingly used for research on machine learning. Rather than comparing means, an SVM creates two separating multi-dimensional hyperplanes (support vectors) to establish a margin between categories. We used the R installation of the libsvm library (Chang and Lin, 2001) in package e1071 (Dimitriadou et al., 2009).

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 2. 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).

Model 1: 82.4%			Model 2: 79.1%		
	predicted			predicted	
true	s	ns	true	s	ns
s	35	5	s	34	7
ns	11	40	ns	12	38
Model 3: 89.0%			Model 4: 92.3%		
	predicted			predicted	
true	s	ns	true	s	ns
s	44	8	s	43	4
ns	2	37	ns	3	41

Table 2. Contingency tables and total accuracies for predictions of different SVM classifiers using OHOCV for binary classification of subject and non-subject conditions.

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* ($4 \cdot (\text{total duration}) / 10$) yielding a best model (Model 4) with 92.3% total accuracy (7 misclassifications).

While it is usually difficult to represent a multi-dimensional SVM graphically, we can get a rough idea by plotting the *s* and *ns* data points as f_1 vs. f_2 at 40% of *I* (Figure 3).

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. What is perhaps more remarkable, however, is that the classifier performed so well with so little information.

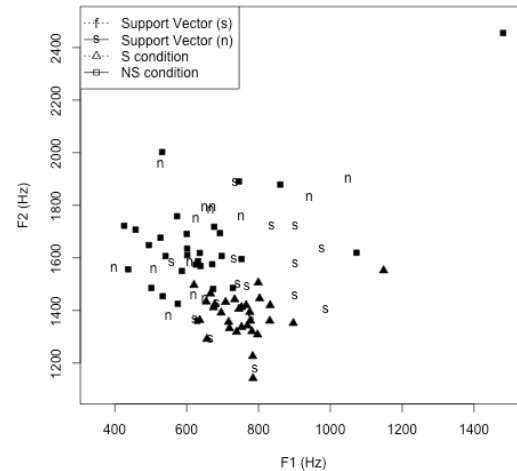


Figure 3. Plot of first formant values against second formant values at 40% location of *I*. Support vectors from Model 4 indicated for *s* and *ns* conditions.

All data in the SVM classifier were scaled by default, allowing a kind of normalization across tokens. However, little or no within-utterance normalization was utilized. Model 4 contained acoustic information (viz. duration) from two different words: *I* and *did*, allowing for limited approximation of rate-of-speed; however, Model 3 contained information from *I* alone.

That formant values are such a strong indicator of focus in the reported classifier (and others omitted here for reasons of space) is consistent with hyperarticulation and featural enhancement models of stress (e.g. de Jong, 1995; Fowler, 1995; Cho, 2005). In broad terms, such models predict that speakers will meet or overreach articulatory targets in stressed syllables, with the effect of greater acoustic/perceptual distinctness, particularly in the first and second formants. For example, Cho (2005) found that a low back vowel tends to be lower and backer, as shown both by tongue position and f_1 and f_2 : a high f_1 is correlated with a low tongue position and a low f_2 is correlated with back tongue position. In our data, the location at 40% of *I* roughly corresponds to the target or extremum of the formants in the first segment [a] (back, round vowel) of the diphthong [aI]: the first formant tends to be higher and the second formant lower for the “s” (subject focus) condition.

Together, the cross-utterance comparison and use of formant extrema favor a model of focus-detection based on paradigmatic enhancement (increased segmental distinction across tokens) over a model based on syntagmatic enhancement, which would instead involve the comparison of non-segmental parameters within a given utterance (Cho, 2005). The classification task is one of evaluating a novel token of *I*, for example, with respect to previously learned tokens, rather than with respect to neighboring words within the utterance.

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 highly successful in the detection of focus placement, we also 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* [ɪ]. 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* described 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 1974; Creswell 2002; König & Gast 2004). 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 *f0*. 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.

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*, *d*, and *f*.

Automation of any of the steps would speed up creating a database. Given a word transcription, there are available solutions for creating a word level alignment. 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.

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 preliminary 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.

A comparable 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 by Calhoun *et al.* (2005), including annotation for syntactic structure as part of the Penn Treebank (Marcus *et al.*, 1993), dialog act (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.*, 1999). 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 yet it is not large enough to include a significant subset of data with specific linguistic constructions such as the comparative “than I did” investigated in our pilot. Annotation by human experimenters also necessarily introduces certain theoretical assumptions such as the prosodic ontology of the TOBI system (Silverman *et al.*, 1992) for prosodic annotation.

The apparent loss of reproducibility with the web search methodology can be compensated for in two ways. First, online public publishing of target files offers a lasting “snapshot” of the larger corpus and record of the particular database created. For example, one might publish short excerpts containing the target *than I did*, provided such publication observes appropriate copyright regulations. Relevant considerations for “fair use” exception under US copyright law are that the purpose is non-commercial research and education; the audio data are used for purposes other than the intended one; the data have

been publicly distributed on the web; the data segments are tiny (as little as one second, if limited to the target string); and that there is no market impact, with the data element not substituting for the original, and little possibility of the copyright holders developing a licensing market for the use to which the element is put.

Second, the dynamic nature of the web means that a generalization may often be tested with a novel set of data. For example, as of January 2006, Everyzing reported its corpus of indexed podcasts to number over 48,797 “and growing” (Lieberman, 2006), and other corpora promise to emerge in the near future thanks to work by organizations such as Google Labs and the MIT Computer Science and Artificial Intelligence Laboratory.

Even without direct reproducibility, however, web corpora are certainly relevant for informal exploration of data, exploratory data analysis, and modeling.

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. *Glott International* 5:341-345.
- Sasha Calhoun. 2008. Why do we accent words? The processing of focus and prosodic Structure. Presented at Experimental and Theoretical Advances in Prosody Conference, Cornell University, NY.
- Sasha Calhoun. 2007. Predicting focus through prominence structure. In *Proceedings of Interspeech 2007*, Antwerp, Belgium.
- Sasha Calhoun. 2006. Information Structure and the Prosodic Structure of English: a Probabilistic Relationship. PhD thesis, University of Edinburgh.
- 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.
- Nick Campbell and Mary Beckman. 1997. Stress, prominence, and spectral tilt. In *Intonation: Theory, models and applications (proceedings of an esca workshop, September 18-20, 1997, Athens, greece)*, ed. by George Carayiannis, Antonis Botinis and Georgios Kouroupetroglou. ESCA and University of Athens Department of Informatics.
- William R. Cantrall. 1973. Why I would relate ‘own’, emphatic reflexives, and intensive pronouns, my own self.’ *Papers from the Ninth Regional Meet-*

- ing, 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>.
- Taehong Cho. 2005. Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /α,i/ in English. *Journal of the Acoustical Society of America* 117:3867-3878.
- Corina Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20:273-297.
- Cassandra 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.
- Carol A. Fowler 1995. Acoustic and kinematic correlates of contrastive stress in spoken English. In *Producing Speech: Contemporary Issues: For Katherine Safford Harris*, eds. F. Bell-Berti and J. J. Raphael, pp. 355-373.
- 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*.
- 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.
- Jochen Puchalla. 2008. Cutmp3. <http://www.puchalla-online.de/cutmp3.html>.
- Ray Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press.
- de Jong, Kenneth J. 1995. The supraglottal articulation of prominence in english: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America* 97:491-504.
- Daniel Jurafsky, Alan Bell, Eric Fosler-Lussier, Cynthia Girand, and William Raymond. 1998. Reduction of English function words in Switchboard. *Proceedings of ICSLP-98*, Volume 7.
- Mark Liberman. Podzinger rejects Jesus. Language Log blog posting, January 25, 2006. URL <http://itre.cis.upenn.edu/~myl/languageelog/archives/002785.html>
- Marcus P. Mitchell, 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.
- R Development Core Team. 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>, ISBN 3-900051-07-0.
- 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*.
- Agaath Sluijter and Vincent van Heuven. 1996. Spectral balance as acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America* 100:2471-2485.
- 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*, 2008.
- Jiahong Yuan and Mark Liberman. 2008b. Vowel Acoustic Space in Continuous Speech: An example of using audio books for research. *CatCod* 2008.