Comparing Automatic Rich Transcription for Portuguese, Spanish and English Broadcast News

Fernando Batista [1,2], Isabel Trancoso [1,3], Nuno J. Mamede [1,3]

[1] *Spoken Language Systems Laboratory, INESC ID Lisboa*
R. Alves Redol, 9, 1000-029 Lisboa, Portugal
{fmmb,imt,njm}@l2f.inesc-id.pt
[2] *DCTI, ISCTE-IUL, Portugal*
Av. Forças Armadas, Lisboa, Portugal
[3] *IST, Technical University of Lisbon*
Av. Rovisco Pais, Lisboa, Portugal

***Abstract*—This paper describes and evaluates a language independent approach for automatically enriching the speech recognition output with punctuation marks and capitalization information. The two tasks are treated as two classification problems, using a maximum entropy modeling approach, which achieves results within state-of-the-art. The language independence of the approach is attested with experiments conducted on Portuguese, Spanish and English Broadcast News corpora. This paper provides the first comparative study between the three languages, concerning these tasks.**

I. Introduction

The text produced by a standard speech recognition system consists of raw single-case words, without punctuation marks, with numbers written as text, and with many different types of disfluencies. The missing information makes this representation format hard to read and understand [1], and pose problems to further automatic processing. Capitalization is important for improving human readability, parsing, and NER (Named Entity Recognition). Several studies have also shown that punctuation marks, or at least sentence boundaries, are important for parsing, information extraction, machine translation, extractive summarization and NER.

These tasks are relevant modules of the Broadcast News (BN) processing system developed at our lab, which integrates several other core technologies, in a pipeline architecture: jingle detection, audio segmentation, automatic speech recognition (ASR), topic segmentation and indexation, and summarization. The first modules of this system, including punctuation and capitalization, were optimized for on-line performance, given their deployment in the fully automatic subtitling system that is running on the main news shows of the public TV channel in Portugal, since 2008 [2]. This BN processing chain was originally developed for European Portuguese, but it was recently ported to Spanish, English, and other varieties of Portuguese (Brazilian and African).

The goal of this paper is to test the language independence of our approach, using the thee languages for which we have recently developed our system; and to compare the proximity of the achieved results with state-of-the-art results.

This paper is organized as follows: the related word and the approach are described in Sections II and III. Section IV presents experimental results for the three languages, concerning automatic punctuation and capitalization. Section V presents the final remarks and the future work.

II. Related work

Whilst the speech-to-text core technologies have been developed for more than 30 years, the metadata extraction/annotation technologies are receiving significant importance only during the latest years. For example, [3] contains an entire section dedicated to this subject, while this topic is only briefly mentioned in the first version of this book, published in 2000. Producing rich transcripts usually involves the process of recovering structural information and the creation of metadata from that information. Recovering punctuation marks and capitalization are two relevant MDA (Metadata Annotation) tasks, which contribute to enriching the final recognition output.

The first joint initiatives concerning automatic rich transcription of speech started around 2002 concomitantly with the DARPA-sponsored EARS program and the NIST RT evaluation series. One of the targets of the five year project EARS program was to advance the state-of-the-art in automatic Rich Transcription of speech. The NIST RT evaluation series[1] is another important initiative that supports some of the goals of the EARS program, providing means to investigate and evaluate STT (speech-to-text) and MDE (Metadata Extraction) technologies, and promote their integration. Nevertheless, despite the emerging efforts for producing rich transcripts, only a few of the most important MDE tasks are covered by these evaluation plans. Most of the current research focus on the English language. However, some initiatives have also been reported for other languages in the last few years. For example, the ESTER evaluation campaign started in 2003 and focus on the evaluation of rich transcription and indexing of BN for the French language [4].

[1] http://www.nist.gov/speech/tests/rt/

Two different rich transcription methods are proposed and evaluated by [5]. The first method consists of adapting the ASR system for dealing with both punctuation and capitalization. This is done by duplicating each vocabulary entry with the possible capitalized forms, modeling the full stop with silence, and training with capitalized and punctuated text. The second method consists of using a ruled-based NE tagger and punctuation generation. The paper shows that the first method produces worse results, due to the distorted and sparser language model. This result suggests the separation of the punctuation and capitalization tasks from the speech recognition system. The rest of this section describes in more detail the previous work related to each one of the tasks.

A. Punctuation

Different punctuation marks can be used in spoken texts, including: *comma*; *period* or *full stop*; *exclamation mark*; *question mark*; *colon*; *semicolon;* and *quotation marks*. However, most of these marks rarely occur and are quite difficult to automatically insert or evaluate. Hence, most studies focus either on *full stop* or in *comma*, which have much higher corpus frequencies. Moreover, previous work on other punctuation marks, such as *question mark* and *exclamation mark*, have not shown promising results [6].

Comma is the most frequent punctuation mark, but it is also the most problematic because it serves many different purposes. It can be used to: introduce a word, phrase or construction; separate long independent constructions; separate words within a sentence; separate elements in a series; separate thousands in a number; and also to prevent misreading. [7] describes a method for inserting *commas* into text, and presents a qualitative evaluation based on the user satisfaction, concluding that the system performance is qualitatively higher than the sentence accuracy rate would indicate.

When dealing with speech, the notion of utterance [3] or sentence-like unit (SU) is often used [8] instead of “sentence”. Detecting an SU, or finding the SU boundaries, roughly corresponds to the task of detecting positions where a punctuation mark is missing. SU boundary detection has gained increasing attention during recent years, and it has been part of the NIST rich transcription evaluations. A general HMM (Hidden Markov Model) framework that allows the combination of lexical and prosodic clues for recovering *full stop*, *comma* and *question marks*, is used by [6] and [9]. A similar approach was also used for detecting sentence boundaries by [10], [11], [12]. [9] also combines 4-gram language models with a CART (Classification and Regression Tree) and concludes that prosodic information highly improve the results. [13] describes a maximum entropy (ME) based method for inserting punctuation marks into spontaneous conversational speech, where the punctuation task is considered as a tagging task and words are tagged with the appropriate punctuation. It covers three punctuation marks: *comma*, *full stop*, and *question mark*; and the best results on the ASR output are achieved using bigram-based features and combining lexical and prosodic features. [14] proposes a multi-pass linear fold algorithm for sentence boundary detection in spontaneous speech, which uses prosodic features, focusing on the relation between sentence boundaries and break indices and duration, covering their local and global structural properties.

B. Capitalization

The capitalization task, also known as truecasing [15], consists of assigning the proper case information to each input word, which may depend on the context. Proper capitalization can be found in many information sources, such as newspaper articles, books, and most of the web pages. Besides improving the readability of texts, capitalization provides important semantic clues for further text processing tasks. The capitalization is not usually considered as a topic by itself. A typical approach, when dealing with processes where capitalization is expected, consists of modifying the process that usually relies on case information in order to suppress the need of that information [16]. An alternate approach is to previously recover the capitalization information, which can also benefit other processes that use case information.

A common approach for capitalization relies on n-gram language models estimated from a corpus with case information [5], [15], [17]. Another approach consists of using a rule-based tagger, as described in [18], which was shown to be robust to speech recognition errors, while producing better results than case sensitive language modeling approaches. [19] describes an approach to the disambiguation of capitalized words where capitalization is expected, such as the first word of the sentence or after a period, which consists of a cascade of different simple positional heuristics. Other approaches include Conditional Random Fields (CRF) and Maximum Entropy Markov Models (MEMM) [20]. A study comparing generative and discriminative approaches can be found in [21]. The impact of using increasing amounts of training data as well as a small amount of adaptation is studied in [20]. Experiments on huge corpora sets using different n-gram orders are performed in [17], concluding that using larger training data sets leads to increasing improvements in performance, but the same tendency is not achieved by using higher n-gram order language models.

III. Approach description

Experiments described in this paper use the same approach for the punctuation and capitalization tasks, which can be treated as two classification tasks. Our experiments use a discriminative approach, based on maximum entropy (ME) models, which provide a clean way of expressing and combining different aspects of the information. This is specially useful for the punctuation task, given the broad set of lexical, acoustic and prosodic features that can be used. This approach requires all information to be expressed in terms of features causing the resultant data file to become several times larger than the original one. On the other hand, the memory required for training with this approach increases with the size of the corpus (number of observations). This constitutes a training problem, making it difficult to use large corpora for training.

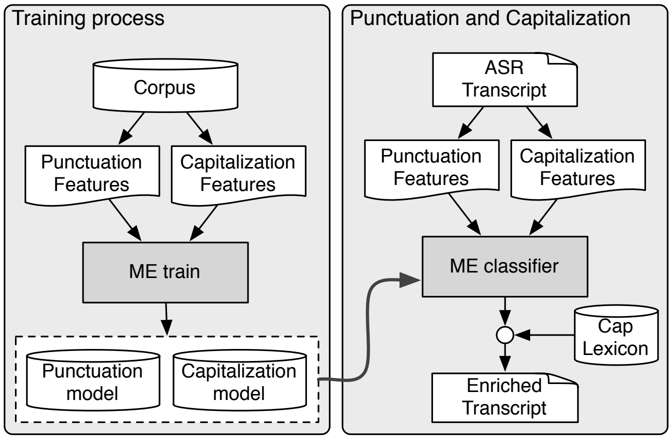

Figure 1. Block diagram of the punctuation and capitalization tasks.

Table I
PORTUGUESE BN CORPUS PROPERTIES.

	#Words	Dur.	*full stops*	*commas*	WER
Train	477k	52h	4.9%	7.8%	11.3%
Devel	66k	7h	4.6%	10.4%	20.8%
Eval	269k	29h	4.6%	5.9%	21.7%

Table II
SPANISH BN CORPUS PROPERTIES.

	#Words	Dur.	*full stops*	*commas*	WER
Train	152k	15h	3.9%	5.9%	11.0%
Devel	25k	3h	4.0%	6.0%	17.2%
Eval	16k	2h	4.5%	4.6%	18.9%

However, the classification is straightforward, making it interesting for on-the-fly usage.

Capitalization models are usually trained using large written corpora, which contain the required capitalization information. The consequent memory problem is solved by splitting the corpus into several subsets, and then iteratively retraining with each one separately. The first subset is used for training the first ME model, which is then used to provide initial weights for the next iteration over the next subset. This process goes on until all subsets are used. Although the final ME model contains information from all corpora subsets, events occurring in the latest training sets gain more importance in the final model. As the training is performed with the new data, the old models are iteratively adjusted to the new data. This approach provides a clean framework for language dynamics adaptation: (1) new events are automatically considered in the new models; and (2) with time, unused events slowly decrease in weight [22], [23].

Figure 1 illustrates the classification approach for both tasks. An updated capitalization lexicon containing the capitalization of new words and mixed-case words can be used as a complement for capitalization. The experiments described in this paper use the `MegaM` tool [24], which uses conjugate gradient and logistic regression.

IV. EXPERIMENTAL RESULTS

This section describes experiments on recovering punctuation marks and capitalization information for Portuguese, Spanish and English. The evaluation is performed using the performance metrics: Precision, Recall and SER (Slot Error Rate) [25]. Only capitalized words (not lowercase) and punctuation marks are considered as slots and used by these metrics. Hence, for example, the SER for the capitalization task is computed by dividing the number of capitalization errors by the number of capitalized words in the reference data.

Tables I, II and III show details of Portuguese, Spanish and English BN corpora subsets, respectively, which were used for training and evaluating our approaches. The Portuguese corpus is a subset of the Broadcast News European Portuguese Corpus, collected during 2000 and 2001. The Spanish BN corpus is a recent corpus, collected during 2008 and 2009. The English BN corpus combines different corpora subsets, available from the Linguistic Data Consortium (LDC). The first 80% of the LDC2005T24 corpus (RT-04 MDE Training Data Text/Annotations) were used for training, 10% for development and the last 10% for evaluation. The Portuguese data contains the highest percentage of *commas*, comparing with the other two languages, and that occurs also in newspaper data. On the other hand, the Portuguese written language contains the lowest percentage of *full stops,* corresponding to longer sentences, but that does not occur in the speech data.

The manual orthographic transcription of these corpora provides the reference data, and includes punctuation marks and capitalization information. For each corpus we had access not only to the manual transcription, but also to the automatic transcription produced by our recognition system [2]. Whereas the manual transcriptions already contain reference punctuation marks and capitalization, this is not the case of the automatic transcriptions. The required reference was produced by means of word alignments between the manual and automatic transcription. The alignment was performed using the NIST SCLite tool[2], followed by an automatic post-

Table III
ENGLISH BN CORPUS PROPERTIES.

	LDC catalog	#Words	Dur.	*full stops*	*commas*	WER
Train	1998T28 2005T24	711k	81h	5.0%	3.4%	32.9%
Devel	1998T28 2005T24	66k	6h	5.4%	4.8%	20.6%
Eval	2000S86 2000S88 2005T24 2007S10	99k	9h	5.1%	4.7%	25.1%

[2] available from http://www.nist.gov/speech.

Table IV
PUNCTUATION MARKS REPLACEMENTS.

Symbol	Replacement
. : ; ! ? ...	*full stop*
, –	*comma*

processing stage, for correcting possible SCLite errors and aligning compound words which can be written/recognized differently. Each corpora subset was automatically annotated with part-of-speech information. The Spanish and the English corpora were annotated using TreeTagger – a language independent part-of-speech tagger [26]. The Portuguese data was annotated using MARv [27].

The higher WER (Word Error Rate) for Portuguese compared to Spanish may be attributed to the larger proportion of spontaneous speech (37% vs. 15%), as well as the higher complexity of the Portuguese phonological system. On the other hand, the English data was recognised using our speech recognition system, with old models, trained with different data, which explains the highest WER.

A. Punctuation

The punctuation experiments here described use data collected from broadcasted TV shows, and consider only the two most frequent punctuation marks: *full stop* and *comma*. All the other punctuation marks were converted into one of these two punctuation marks, in accordance with the replacements described in Table IV.

The following features were used for a given word w in the position i of the corpus: w_i, w_{i+1}, $2w_{i-2}$, $2w_{i-1}$, $2w_i$, $2w_{i+1}$, $3w_{i-2}$, $3w_{i-1}$, p_i, p_{i+1}, $2p_{i-2}$, $2p_{i-1}$, $2p_i$, $2p_{i+1}$, $3p_{i-2}$, $3p_{i-1}$, $GenderChgs_1$, $SpeakerChgs_1$, and $TimeGap_1$, where: w_i is the current word, w_{i+1} is the word that follows and $nw_{i\pm x}$ is the n-gram of words that starts x positions after or before the position i; p_i is part-of-speech of the current word, and $np_{i\pm x}$ is the n-gram of part-of-speech of words that starts x positions after or before the position i. $GenderChgs_1$, and $SpeakerChgs_1$ correspond to changes in speaker gender, and speaker clusters; $TimeGap_1$ corresponds to the time period between the current and following word. For the moment, only standard lexical and acoustic features are being used in this task. Nevertheless, prosodic features, which already proved useful for this task, will be included in future experiments.

Tables V, VI and VII show the results achieved for Portuguese, Spanish and English data, respectively. The overall results are affected by the *comma* detection performance, which is significantly lower, in terms of SER. The higher SER values for the *comma* for Portuguese result from the data being annotated by different people in different time periods, using possibly different criteria. The precision values are consistently better than the recall values, which reveals that the system usually prefers to avoid mistakes than to add incorrect slots. As expected, the planned speech achieves better performances than spontaneous speech, which significantly affects the overall performance on the Portuguese data, due to the large portion of spontaneous speech in the corpus. Results are strongly affected by the presence of recognition errors, as shown in the performance difference between manual and automatic transcripts. Despite the WER in the English data being the highest, the same conclusion does not equally apply to this data, where results are almost similar for manual and automatic transcripts. That may be due to the presence of additional information in the automatic transcripts, specially on pause duration.

The work reported in [6] also uses the 1998 Hub-4 evaluation data (LDC2006S86) for testing, and reports a set of experiments, using finite state models that use either *pause duration* or *phone duration*. The best performance reported by the paper, when all punctuation marks are combined, is 89% SER, while our experiments achieve a SER 16% better (excluding the *question mark*). The paper also reports results on individual punctuation marks, achieving 41% to 79% SER for the *full stop*, and 81% to 110% SER for the *comma*, but these results are insufficient for drawing conclusions. The 1998 Hub-4 evaluation data was also used in [9], but only a portion of that data was used for evaluation purposes, so results cannot be directly compared. These two studies, make use of acoustic, lexical and prosodic features for recovering *full stop*, *comma* and *question mark*; and [9] concludes that, when prosodic information is used, F-measure can be improved by 19% relative. Despite evaluation sets being different and the question mark not being included in our experiments, our results are similar to the results reported by [9] for automatic transcripts, working with a 17% recognition WER (our system achieves 26% WER for the LDC2000S86 test set).

The recent study described in [17] considers different language models, using from 58 million to 55 billion words for training, and 3-grams to 6-grams. The work reports 52% F-measure (48% precision, 56% recall) using 55 billion words for training, and a significant decrease in performance for the smaller training set (45% precision, 32% recall). Our experiments use less than one million words of speech transcripts, which suggests that our results can be much improved by using bigger training sets of data. Their best result concerning the *full stop* detection (62.5% F-measure) is similar to our result for English data (63.5% F-measure).

The SER, now widely accepted for this task, is strongly influenced by substitutions between punctuation marks. However, [6] argues that replacing a comma by a full stop or vice versa would still aid structuring the output, and would be better than having no punctuation marks at all. Concerning this subject, [9] chooses to count substitutions as half an error, resulting in a decreased SER. Using such scoring strategy our SER results on the English data improve by 4% to 5%.

B. Capitalization

The capitalization experiments assume that the first word of each sentence is processed in a separated processing stage (e.g. after punctuation), since its correct graphical form depends on

Table V
PUNCTUATION RESULTS FOR THE PORTUGUESE BN CORPUS.

	Manual Transcripts									Automatic Transcripts								
Focus	Full stop			Comma			ALL			Full stop			Comma			ALL		
	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER
All	82.9	69.4	44.9	42.3	28.2	110.2	62.9	46.8	65.9	64.8	64.4	70.5	30.0	22.7	130.3	48.0	41.5	88.9
Planned	86.5	71.1	40.0	36.8	24.7	117.6	66.7	50.3	57.5	72.2	70.1	56.9	27.6	25.0	140.6	52.8	49.8	79.1
Spontaneous	77.0	65.5	54.0	46.3	31.4	105.0	58.3	43.0	75.3	51.6	52.8	96.7	32.3	21.4	123.5	40.8	32.0	100.8

Table VI
PUNCTUATION RESULTS FOR THE SPANISH BN CORPUS.

	Manual Transcripts									Automatic Transcripts								
Focus	Full stop			Comma			ALL			Full stop			Comma			ALL		
	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER
All	87.0	67.4	42.7	51.2	31.3	98.5	71.4	49.5	61.7	76.9	58.9	58.9	43.5	23.5	107.0	63.1	41.2	73.9
Planned	87.9	67.3	41.9	52.8	31.4	96.6	72.8	49.5	59.5	82.3	58.1	54.3	48.0	23.5	102.0	68.3	40.9	69.0
Spontaneous	85.3	71.8	40.5	49.4	28.9	100.7	69.9	49.5	66.2	68.0	62.2	67.1	32.7	20.2	121.3	53.0	40.3	86.9

Table VII
PUNCTUATION RESULTS FOR THE ENGLISH BN CORPUS.

	Manual Transcripts									Automatic Transcripts								
subset	Full stop			Comma			ALL			Full stop			Comma			ALL		
	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER	Prec	Rec.	SER
All	73.1	56.2	64.5	62.2	13.5	94.7	70.9	36.1	69.4	72.5	58.8	63.5	63.5	7.0	97.0	71.5	34.4	72.2
LDC2000S86	69.3	52.8	70.6	55.0	12.0	97.8	66.5	34.1	73.3	66.8	56.3	71.7	51.3	5.4	99.7	65.3	32.8	76.7
LDC2000S88	73.5	50.8	67.5	54.5	12.4	98.0	69.7	34.1	72.1	71.1	56.5	66.5	51.8	5.9	99.6	69.2	34.4	74.2

its position in the sentence. Only three ways of writing a word will be considered: lower-case, first-capitalized, and all-upper. Mixed-case words, such as "McGyver", are being dealt with by means of a small lexicon, but are not evaluated in the scope of this paper.

The following features were used for a given word w in the position i of the corpus: w_i, $2w_{i-1}$, $2w_i$, $3w_{i-2}$, $3w_{i-1}$, where w_i is the current word, w_{i+1} is the word that follows and $nw_{i\pm x}$ is the n-gram of words that starts x positions after or before the position i.

The Portuguese capitalization model was trained with a newspaper corpus collected from 1999 to 2004 and containing about 148M words. The Spanish capitalization model was trained with the content of online text, daily collected since 2003, and containing about 79M words. The English capitalization model was trained using content of North American News Text Supplement (LDC1998T30), collected between 1994 and 1998, and containing about 461M words. The original texts were normalized and all the punctuation marks removed, making them close to speech transcriptions. Only data previous to the evaluation data period was used for training.

The retraining approach described in Section III was followed, and the most recent capitalization model was used for processing each evaluation subset. Table VIII shows the corresponding results for each one of the languages.

Table VIII
CAPITALIZATION RESULTS.

Corpus	Written Corpora			Manual Transc.			Automatic Transc.		
	Prec	Rec	SER	Prec	Rec	SER	Prec	Rec	SER
Portuguese	93.8	86.5	19.0	84.4	86.7	29.1	73.2	77.7	50.4
Spanish	95.1	83.2	20.8	94.7	85.6	19.0	77.6	74.1	47.1
English	95.8	81.8	21.1	89.2	80.5	28.9	73.9	71.5	53.2

While the achieved performance is similar for written corpora in all languages, there is a significant difference for the speech data. Three different factors have contributed to the best results achieved for the Spanish language: the first and most important factor is that the manual transcripts were produced by using data previously capitalized by our system, where the annotator simple corrected what was wrong; the second reason is that the bigger portion of planned speech in compared to Portuguese; finally the WER on this corpus is the lowest, which contributes to the best results on the automatic transcripts.

Results concerning English and Portuguese BN data are similar. While the English language achieves a better performance for the manual transcripts, the Portuguese language achieves better performance on the automatic transcripts. The worse performance for the Portuguese data may be due to the unusual topic covered in the news by that time (War on Terrorism). Many foreign names, which can be rarely found in the news, were used by that time. The worse performance for English automatic transcripts is correlated with the higher WER that characterizes this data.

These results on capitalization are difficult to compare to other related work, mainly because of the different evaluation sets, but also because of the different evaluation metrics and applied criteria. For example, sometimes it is not clear whether the evaluation takes into consideration the first word of each sentence. However, these results are consistent with the recent work reported by [17], which achieves 88.5% F-measure (89% precision, 88% recall) on written corpora (WSJ) and 83% F-measure (83% precision, 83% recall) for manual transcripts. Our experiments achieve 88% F-measure for written corpora and 85% F-measure for manual transcripts.

V. Conclusions

This paper presents a language independent approach for recovering punctuation marks and capitalization information over speech data, and describes experiments conducted over Portuguese, Spanish and English, attesting the language independence of the approach. Achieved results are within state-of-the-art. The described approach is now implemented by two modules, one for recovering punctuation marks and the other for recovering capitalization information. These modules have been integrated in our speech recognition system, currently being used to daily process BN shows on-the-fly, for automatic subtitling.

We plan to port the punctuation and capitalization modules to other varieties of the Portuguese Language, for which we recently developed our ASR system, such as Brazilian Portuguese. Porting the modules to new languages will also be an important issue in the future. We are currently trying to further improve the performance of the punctuation module by introducing prosodic features, besides the current lexical and acoustic features. The study on punctuation marks will be extended in order to cover the *question mark*, due to its importance for the text readability. We also plan to further research the low performance of comma insertion, by studying the agreement in different human annotations.

Acknowledgment

The authors would like to thank Raquel Martinez and Thomas Pellegrini for their support with the Spanish and English corpora, and Hugo Meinedo for the support on the speech recognition. This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008 and supported by DCTI, ISCTE-IUL – Lisbon University Institute. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

References

[1] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedorenko, D. Reynolds, and M. Zissman, "Measuring the readability of automatic speech-to-text transcripts," in *Proc. of Eurospeech*, 2003, pp. 1585–1588.

[2] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. of ICASSP 2008*, 2008, pp. 1561–1564.

[3] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall, 2008, ch. 10 - Speech Recognition: Advanced Topics.

[4] G. Gravier, J.-F. Bonastre, E. Geoffrois, S. Galliano, K. M. Tait, and K. Choukri, "The ESTER evaluation campaign for the rich transcription of french broadcast news," in *Proc. LREC 2004*, 2004.

[5] J.-H. Kim and P. Woodland, "Automatic capitalisation generation for speech input," *Computer Speech & Language*, vol. 18, no. 1, pp. 67–90, 2004.

[6] H. Christensen, Y. Gotoh, and S. Renals, "Punctuation annotation using statistical prosody models," in *Proc. of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001, pp. 35–40.

[7] D. Beeferman, A. Berger, and J. Lafferty, "Cyberpunc: a lightweight punctuation annotation system for speech," *Proc. of the ICASSP-98*, pp. 689–692, 1998.

[8] S. Strassel, *Simple Metadata Annotation Specification V6.2*, online, Linguistic Data Consortium, 2004.

[9] J. Kim and P. C. Woodland, "The use of prosody in a combined system for punctuation generation and speech recognition," in *Proc. Eurospeech*, 2001, pp. 2757–2760.

[10] Y. Gotoh and S. Renals, "Sentence boundary detection in broadcast speech transcripts," in *Proc. of the ISCA Workshop: ASR-2000*, 2000, pp. 228–235.

[11] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür, "Prosody-based automatic segmentation of speech into sentences and topics," *Speech Communications*, vol. 32, no. 1-2, pp. 127–154, 2000.

[12] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[13] J. Huang and G. Zweig, "Maximum entropy model for punctuation annotation from speech," in *Proc. of the ICSLP*, 2002, pp. 917 – 920.

[14] D. Wang and S. S. Narayanan, "A multi-pass linear fold algorithm for sentence boundary detection using prosodic cues," in *Proc. ICASSP 2004*, vol. 1, 2004, pp. 525–528.

[15] L. V. Lita, A. Ittycheriah, S. Roukos, and N. Kambhatla, "tRuEcasIng," in *Proc. of ACL-03*, 2003, pp. 152–159.

[16] E. Brown and A. Coden, "Capitalization recovery for text," *Information Retrieval Techniques for Speech Applications*, pp. 11–22, 2002.

[17] A. Gravano, M. Jansche, and M. Bacchiani, "Restoring punctuation and capitalization in transcribed speech," in *ICASSP 2009*, Taipei, Taiwan, 2009.

[18] E. Brill, "Some advances in transformation-based part of speech tagging," in *AAAI '94: Proc. of the 12^{th} national conference on Artificial intelligence*, vol. 1, 1994, pp. 722–727.

[19] A. Mikheev, "A knowledge-free method for capitalized word disambiguation," in *Proc. of the ACL-99*, 1999, pp. 159–166.

[20] C. Chelba and A. Acero, "Adaptation of maximum entropy capitalizer: Little data can help a lot," *Proc. of the EMNLP '04*, 2004.

[21] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[22] F. Batista, N. Mamede, and I. Trancoso, "Language dynamics and capitalization using maximum entropy," in *Proc. of ACL-08: HTL - Short Papers*, 2008, pp. 1–4. [Online]. Available: http://www.aclweb.org/anthology/P/P08/P08-2001

[23] F. Batista, N. Mamede, and I. Trancoso, "The impact of language dynamics on the capitalization of broadcast news," in *Proc. of Interspeech 2008*, Sep. 2008.

[24] H. Daumé III, "Notes on CG and LM-BFGS optimization of logistic regression," 2004, http://hal3.name/megam/.

[25] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of the DARPA BN Workshop*, 1999.

[26] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom, 1994.

[27] R. Ribeiro, L. Oliveira, and I. Trancoso, "Using morphossyntactic information in TTS systems: comparing strategies for european portuguese," in *Proc. of PROPOR 2003*. Springer, 2003, pp. 26–27.