Revising the annotation of a Broadcast News corpus: a linguistic approach

V. Cabarrão[1,2], H. Moniz[1,2], F. Batista[1,3], R. Ribeiro[1,3], N. Mamede[1,4], H. Meinedo[1], I. Trancoso[1,4], A. I. Mata[2], D. M. de Matos[1,4]

[1] L2F, INESC-ID
[2] FLUL/CLUL - Universidade de Lisboa
[3] ISCTE - Instituto Universitário de Lisboa
[4] IST - Universidade de Lisboa
veracabarrao@gmail.com, {Helena.Moniz, Fernando.Batista, ricardo.ribeiro, Nuno.Mamede, Hugo.Meinedo, Isabel.Trancoso}@inesc-id.pt, aim@fl.ul.pt, david.matos@l2f.inesc-id.pt

Abstract

This paper presents a linguistic revision process of a speech corpus of Portuguese broadcast news focusing on metadata annotation for rich transcription, and reports on the impact of the new data on the performance for several modules. The main focus of the revision process consisted on annotating and revising structural metadata events, such as disfluencies and punctuation marks. The resultant revised data is now being extensively used, and was of extreme importance for improving the performance of several modules, especially the punctuation and capitalization modules, but also the speech recognition system, and all the subsequent modules. The resultant data has also been recently used in disfluency studies across domains.

Keywords: Speech annotation, Metadata, Broadcast News

1. Introduction

Speech-to-text core technologies have been developed over more than 30 years, but the text produced by a standard Automatic Recognition System (ASR) consists of raw single-case words without structural metadata segmentation, such as punctuation marks, and without any knowledge about disfluencies. Such representation is hard to read (Jones, 2005) and poses problems to further automatic processing, due to the missing structural metadata. For that reason, metadata extraction/annotation technologies became recently of significant importance. The need to produce rich transcriptions motivated the revision of our most representative and most relevant corpus for model training. Such structural metadata events (Liu *et al.*, 2006; Ostendorf *et al.*, 2008) were absent or non-uniformed in previous annotations of the corpus. The revised annotation aimed at providing a suitable sample for further spoken language processing analysis in the broadcast news domain, focusing on rich transcriptions. Underlying the goal of this revision process is the fact that linguistic annotation of corpora is the backbone of supervised methods for statistical natural and spoken language processing. Therefore, the annotation and unification of a corpus is a crucial process. This is even more salient when the data is the input of a pipeline architecture, cascading possible errors into different modules of this architecture.

2. Corpus description

The focus of the work here reported is the ALERT corpus, an European Portuguese Broadcast News corpus, originally collected from the public TV channel (RTP) for training and testing the speech recognition and topic detection systems, in the scope of the ALERT European project (Neto *et al.,* 2003; Meinedo *et al.,* 2003). It includes 122 programs of different types and was split into different subsets, used for different purposes within the ASR system.

Usage	Name	Recording period	Dur.	Words
Train		Oct., Nov 2000	46.5h	480k
Devel		Dec. 2000	6.5h	67k
Test	Eval	Jan. 2001	4.5h	48k
	JEval	Oct. 2001	13.5h	136k
	RTP07	May, June, Sep., Oct., 2007	4.8h	49k
	RTP08	June, July, 2008	3.7h	39k

Table 1 - The ALERT corpus subsets

Table 1 presents details of each part of the corpus, where duration values represent useful time (silences not included). The training data has around 47h of useful speech encompassing 480k words. The original corpus included two different evaluation sets: Eval and JEval, the latter having been collected with the purpose of a "joint evaluation" among several partners. The evaluation set was later complemented with two collections of 11 BN shows from the same public TV channel (RTP). Our test set combines the four previously mentioned sets, totaling about 26h of data (Meinedo, 2008; Batista, 2011).

3. Review process

The review process of the corpus was motivated by the need to tackle rich transcriptions, especially structural metadata events (Liu *et al.*, 2006; Ostendorf *et al.*, 2008), *i. e*., disfluencies and punctuation marks, that were absent or non-uniformed in previous annotations. The revision also encompassed orthographic, morphosyntactic, and paralinguistic information as well (laughs, coughs, etc.).

Symbols	Context of use	Examples
< >	Auto-corrected sequences	sequences of disfluencies
[ ]	Non analyzable speech sequences	noisy conditions, *inter alia*
& &	delimits onomatopoeic words	&quá quá quá& (the sound made by a duck)
^	Proper names	^António
~	In the right edge of the word stands for irregular pronunciation;	pode-nos <servir~> servir (can <serve~> serve us) pronounced as S@r'nir instead of s@r'vir
~	in the left edge stands for spelled sigla or mathematical expression/variable	~GNR matriz ~A (matrix ~A)
@	Acronyms	@INESC
+	Word contractions or syncopated forms	+está (is) pronounced as ['ta], instead of [S'ta], transcribed in the citation form.
§	Morphosyntactic irregular forms	depois parte destas contas §têm que ser §saldadas (afterwards part of these accounts §they will have to be settled)
%	Filled pauses	%aa (%uh)
--	Word fragment	<comp-> complementar (additional)
=	Excessive segmental prolongations	que= (that=) pronounced as [k@:]

Table 2 - Inventory of diacritics used in the annotation. Phonetic symbols are presented in SAMPA.

3.1 Linguistic annotation

Automatic transcripts provided by the in-house speech recognizer (Neto *et al.*, 2008) are used as a basis that the transcribers correct. At this stage, speech is segmented into chunks delimited by silent pauses, already containing audio segmentation related to speaker and gender identification and background conditions. Each segment in the corpus is marked as: planned speech with or without noise (F40/F0); spontaneous speech with or without noise (F41/F1); telephone speech (F2); speech mixed with music (F3); non-native speaker (F5); any other speech (FX). Most of the corpus consists of planned speech, but it also contains a large percentage (35%) of spontaneous speech. All segmented speech chunks are manually annotated with the set of diacritics presented in Table 2. Both previous and the revised annotations were done using Transcriber.

Disfluencies are important structural metadata events used for online editing spontaneous speech. Filled pauses, segmental prolongations, repetitions, substitutions, deletions, insertions, editing expressions, word fragments, and mispronunciations are annotated following Moniz (2006), based on Shriberg (1994) and Eklund (2004), and basically using the same set of labels. Disfluencies are delimited by angular brackets and are further specified in a separate tier regarding its specific type and structure (for further details see Moniz, 2013).

Speech units do not always correspond to sentences, as established in the written sense. They may be quite flexible, elliptic, restructured, and even incomplete (Blaauw, 1995). As for punctuation marks, we used the guidelines for European Portuguese presented by Duarte (2000), in which punctuation marks are described based on syntactic criteria. Comma deserves a special attention due to its multifunctionality. Thus, the presence/absence of a comma in specific locations may influence the grammatical judgments of speech units. As reported by Duarte (2000), commas should not be placed between: i) the subject and the predicate; ii) the verb and the arguments; iii) the antecedent and the restrictive relative clause; and iv) before the copulative conjunction e/and in an enumeration. However, commas should separate: i) adverbial subordinate clauses; ii) appositive modifiers; iii) parenthetical constituents; iv) anteposed constituents; v) asyndetically coordinated constituents; and vi) vocatives. Moreover, disfluent sequences in itself are not punctuated, because its meaning is often hard to reconstruct, and may correspond to ungrammatical sequences.

Since broadcast news is a domain with very productive multilingual scenarios, annotation criteria must also encompass the identification of other languages. To do so, there are two distinct processes: (i) foreign-origin words that have already been assimilated by the language need not any specific tagging (*e.g.* Internet); (ii) whereas all other cases need the proper language identification. Moreover, times, dates and numerals (cardinals and ordinals) must always be spelled out in regular words, to avoid multiple criteria in the orthographic annotation.

Further criteria also regards spontaneous speech phenomena, such as interjections (*Ah!*, *Oh!*, *Ui!*), grunts (*hum*, *humhum*, *hã*, *hãhã*), backchannels (*ok*, *sim*), and discourse markers (*portanto*, *pronto*), which are still understudied in our language.

3.2 Statistical differences for structural metadata events

The differences between the previous version and the revised data, in terms of the structural metadata events, are summarized in Table 3. For example, about 41k

Punctuation before revision

Punctuation after revision	none	.	,	?	!	...	:	;	"	<>
none		1574	31429	52	67	249	11	6	202	0
.	1239	40064	1856	33	172	63	14	5	0	0
,	18828	3892	41435	83	74	36	62	27	16	0
?	118	233	66	2017	3	11	2	1	0	0
!	14	13	18	0	64	0	0	0	0	0
...	139	30	43	1	1	172	0	0	0	0
:	110	83	140	0	4	0	104	8	1	0
;	63	52	54	0	0	0	0	11	0	0
"	0	0	0	0	0	0	0	0	0	0
<>	14576	6	36	0	0	7	0	0	3	0

>20k
10k-20k
5k-10k
1k-5k
100-1k
20-100
5-20

Table 3 - Confusion matrix of structural metadata events between corpora versions.

commas were kept from the previous to the new version, but about 31k were removed, and another 19k were simply added to the new corpora version. Striking differences with commas are due to the fact that in the previous version annotators used commas to delimit sequences of disfluencies or whenever there was a silent pause or a paralinguistic event, even if the placement of a comma did not respect the syntactic structure (e.g., often introducing a comma between the subject and the predicate as in *A minha presença aqui, tem apenas o sentido de uma homenagem, sentida e, %aa, solidariedade para com as famílias/ My presence here, has only the sense of a felt, tribute and, %eh, solidarity with the families)*.

The bottom line of the table reports statistics for the newly added disfluency boundaries, revealing that about 15k disfluencies are now marked in the corpus. As for full-stops, the table shows that, whereas about 40k full-stops were kept from the previous version to the newer, about 3.9k were replaced by commas, 233 were replaced by question marks, and another 1574, were simply removed. The question mark is the most consistent punctuation mark across annotations. Results concerning other punctuation marks are less significant given the lower frequency in the corpus. In the first version of the annotation, exclamation marks were used in contexts, such as prepared speech produced by the anchor, which are not truly exclamations. Most of them were capturing spontaneous speech for interjections and strong reactions from the interviewees.

Other differences in the tables concern the use of *double quotes*. Being used for citations, they are not very frequent in broadcast news, and they are also quite difficult to deal with using computational scripts. Moreover, in the first version of the annotation they were mostly ascribed to foreign language segments, which are also being identified using specific language identification labels. Therefore, *double quotes* were removed in the second version of the annotations.

3.3 Inter-annotator agreement

Using the differences previously presented, Cohen's kappa values (Carletta, 1996) have been calculated for each one of the structural metadata events, allowing to assess the user agreement between the previous version, transcribed by three annotators, and the final version, revised by an expert linguist. Table 4 shows the corresponding results, revealing that the most consistent punctuation marks are the full-stop and the question-mark, and confirming the strong disagreement concerning comma, in line with the confusion matrix presented in the previous section.

	Kohen's Kappa
full-stop (.)	89.0%
comma (,)	55.7%
question mark (?)	87.0%
exclamation mark (!)	25.9%
...	37.2%
colon (:)	32.3%
semicolon (;)	9.2%
All punctuation marks	70.5%

Table 4 - user agreement between the first and second versions of the transcript corpus.

Considerable differences beyond structural metadata events are also found between the first and second versions of the transcript corpus, mostly due to other idiosyncratic properties of spontaneous speech. European Portuguese is a language with strong reduction and co-articulation effects. These effects, in extreme reduction scenarios, are very hard to transcribe. Such examples are marked with the diacritic "+" before all the contracted forms, instead of just the first word of the sequence. Ascribing the diacritic to the sequence of words subject to such reduction and co-articulation effects allow us to recover the entire sequence (sometimes more than 3 words, +com a azul [kazul~] /with the blue, replaced by +com +a +azul) and, thus, to evaluate the WER (Word Error Rate) of the ASR output in such regions. Another example is non lexicalized forms mistaken with disfluent fragments, such as ['tar] for estar/to be and ['tãw] for estão/are, transcribed in its citation form and not in its

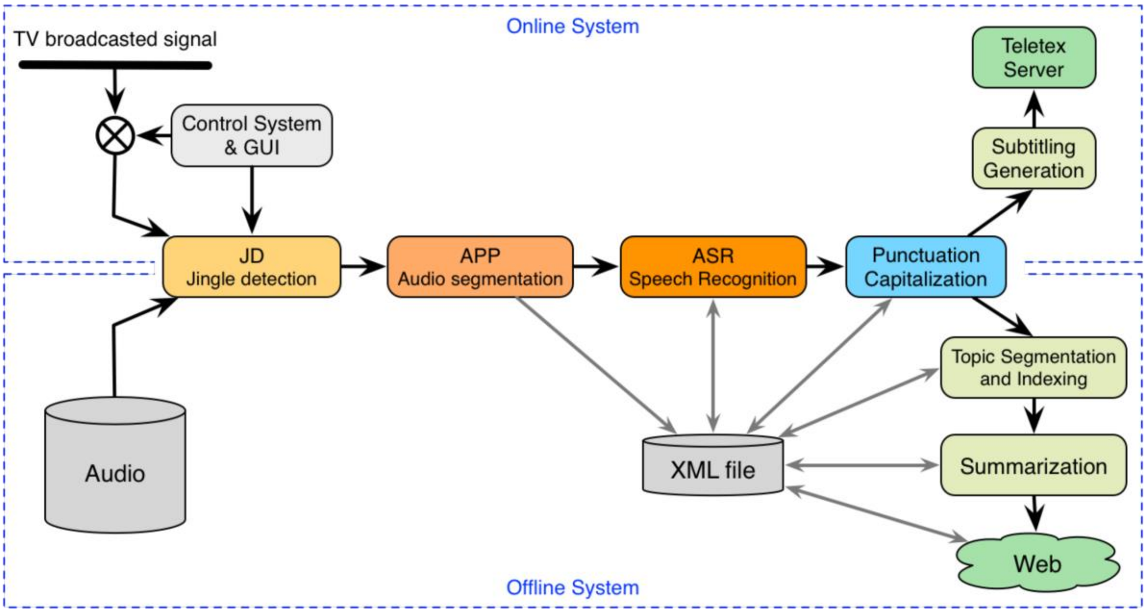

Figure 1 – Automatic Speech Recognition modules.

syncopated form (eu só lhe pergunto a ele, +onde é que –tão (replaced by +estão), os professores que –tão (replaced by +estão) desempregados/ I just ask him, where are the, teachers that are unemployed).
Other problems found concern orthographic and accentuation errors (confusion between vêem/to see and vêm/to come), as well as inappropriate use of uppercase and lowercase in EP (days of the week, professions, etc.), previous to the Orthographic Agreement now in use.

3.4 Automatic processing

The revised data is now a valuable resource for studying and analyzing metadata, and since then a considerable number of studies have been performed (*e.g.*, Meinedo, 2008; Abad & Neto 2008; Meinedo *et al.,* 2010; Batista *et al.,* 2010; Moniz *et al.,* 2010; Moniz *et al.,* 2011; Batista, 2011; Batista *et al.*, 2012a; 2012b; Moniz, 2013; Ribeiro and de Matos, 2013). Currently, the data is being automatically annotated with part-of-speech (POS) tags, initially by MARv (Ribeiro *et al.*, 2003) and more recently by Falaposta (Batista *et al.*, 2012a), a CRF-based tagger robust to certain recognition errors, given that a recognition error may not affect all its input features.
After the speech recognition, all relevant manual annotations, including all metadata events, are being transferred to the ASR transcripts, and constitute extended transcripts, available as self-contained XML files (Batista *et. al.*, 2012a). Each XML corresponds to a transcript integrating both manual and automatic synchronized transcripts, enriched with additional prosodic information related to pitch, energy, duration, and other structural metadata (punctuation, disfluencies, paralinguistic annotation, etc.). Our most recent experiments use a fixed set of purely automatic features, extracted or calculated from the extended transcripts.

4. Impact on different tasks

The in-house ASR system is composed of several modules, as illustrated in Figure 1. This section presents the impact of the revised data on three of these modules, namely: speech recognition (ASR), Punctuation and Capitalization (rich transcription), and speech Summarization.

4.1 Speech recognition

The speech recognition system is a pipeline of modules that change often, which makes it difficult to measure just the impact of building new models. Taking that into account, Table 5 shows the results, in terms of WER, of two evaluations performed on the ASR system, before and after the revised data being integrated in the models. During the revision process, a significant number of orthographic information and corresponding segmentation were corrected. Such modifications turned out to be positive for the performance of the speech recognition models. Interestingly enough, the performance decrease for the training data suggests that the new model might be less prone to overfit.

WER	Train	Devel	Test
Initial annotations	11.3%	20.8%	21.7%
Revised data	14.1%	19.6%	19.4%

Table 5 - Impact of the revised data in the ASR performance.

4.2 Rich transcription

The initial studies on automatic metadata, performed immediately after the revision of the corpus, focused on detecting full-stops and commas, the two most common punctuation marks. Table 5 compares the results achieved with the two versions of the annotations, over the ASR output. By that time, only lexical and time-related features were used for the experiments. The overall results are affected by the lower performance achieved for the comma, in terms of SER (Slot Error Rate), for which the user agreement is also very low (see Table 4). The revised data conducts to a significant increase of performance, especially in terms of commas.

	Initial annotations			Revised data		
	Prec	Rec.	SER	Prec	Rec.	SER
Full stop	64.8	64.4	70.5	68.6	61.6	66.6
Comma	30.0	22.7	130.3	59.5	29.0	90.8
ALL	48.0	41.5	88.9	64.7	42.6	69.8

Table 6 – Overall results between the first and second versions of the transcript corpus.

4.3 Speech summarization

The impact of the revised annotation was also evaluated in speech summarization. Ribeiro and de Matos (2013) propose the use of additional related information sources in the speech summarization. The approach, genre-independent, is assessed over broadcast news. The importance of the manual annotation of the corpus is related to the selection process of the additional information sources. The idea is to select passages from written newspaper articles related to the news story to be summarized. This selection is done by computing a phonetic distance between the sentences of the newspaper article and the automatically recognized sentence-like units from the target news story. If this distance is below an estimated threshold, the sentence from the written article is selected to be included in the summarization process. To estimate the threshold, it is computed the distance between the phonetic transliterations of both the automatic transcription and the manual transcription: we conjecture that the distance between an article from a written newspaper and the automatic transcription of a broadcast news story about the same topic is similar to the distance between the automatic transcription of a news story and its corresponding manual transcription. A simple Latent Semantic Analysis-based (Landauer and Dumais, 1997) extractive summarizer was used as baseline. The analysis of the indirect impact through improved models for speech recognition and rich transcription shows several differences in the generated summaries and some changes in the human judges' preferences. Table 7 shows the WER and SER for the data used for summarization.

	WER	SER
Initial annotations	19.5%	90.2%
Revised data	16.5%	81.5%

Table 7 – Characterization of the data used in the speech summarization experiments.

An interesting difference is that although the ranking in terms of users' preference did not change—human summaries on top, then our approach, and, in the end, the baseline—, the improvements in terms of WER and SER led to an increase of preference for the baseline system. More important were the effects on the composition of the summaries generated by the proposed summarization approach. The additional information was used in two ways: summaries composed only by the additional information and summaries that mixed the content of the original transcribed news story with the additional information. Before the revision, for about 70% of the news stories these summaries were equal; after the revision, they were equal only for 35% of the news stories.

5. Conclusion

This paper describes the revision process of a broadcast news corpus that follows a linguistic approach. This process was made by a linguist with the main concern of standardize the annotation criteria for structural metadata events.

The validity of using very rich corpora with well-structured linguistic descriptions *vs.* big data with no annotation is quite arguable nowadays. This study aims at contributing to this debate, providing evidences that well-structured data, with uniform criteria, improve natural and spoken language processing tasks. This is even more important when studying a language such as Portuguese, since much research is still needed when compared to other languages.

The annotation schema used in this paper proved to be a valid source of information for rich transcription modules in our in-house speech recognizer. The rich transcription modules trained with the revised data are also integrated in SPA (SPeech Analytics), a web service that enables a client (user) to upload an audio-visual document in order to automatically process it according to some desired customizable options. Then, the SPA service produces an automatic Rich Transcription (RT) result that is presented in a friendly and editable transcription web-based framework. In this way, the user can further correct the automatic transcription and corresponding metadata before generating a final result. This web-based service is also an answer to many requests from individuals and companies with transcription needs, for whom a unified framework that delivers a (partially erroneous) RT, and allows edition. Access to the SPA service (https://spa.l2f.inesc-id.pt) is protected but can be granted to individual users upon request. If you would like to try the current implementation and explore its possibilities please send an email to info@l2f.inesc-id.pt describing briefly your intentions.

Additionally, the annotation schema was recently applied under the auspices of the European Project EU FP7, *Collaborative information, Acquisition, Processing, Exploitation and Reporting for the prevention of organized crime* (CAPER), in cooperation with VoiceInteraction.

6. Acknowledgements

This work was supported by national funds through – Fundação para a Ciência e a Tecnologia, under projects PEst-OE/EEI/LA0021/2013 and by ISCTE-IUL, Instituto Universitário de Lisboa.

7. References

Abad, A. and Neto, J. (2008) Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer. In *Interspeech 2008*, Brisbane, Australia.

Batista, F., Moniz, H., Trancoso, I., Meinedo, H., Mata, A. I., Mamede, N., (2010) Extending the punctuation module for European Portuguese, In Proc. of Interspeech 2010, Mukari, Japan, September 2010.

Batista, F. (2011) Recovering Capitalization and Punctuation Marks on Speech Transcriptions. PhD thesis, Instituto Superior Técnico.

Batista, F., Moniz, H., Trancoso, I., Mamede, N., and Mata, A.I. (2012a) Extending automatic transcripts in a unified data representation towards a prosodic-based metadata annotation and evaluation. In *Journal of Speech Sciences*, (3).

Batista, F., Moniz, H., Trancoso, I. and Mamede, N. (2012b) Bilingual Experiements on Automatic Recovery of Capitalization and Punctuation of Automatic Speech Transcripts. In *IEEE Transactions on Audio, Speech and Language Processing*, volume 20, pp. 474-485.

Blaauw, E. (1995) *On the Perceptual Classification of Spontaneous and Read Speech.* PhD thesis. University of Utrecht. Published by LEd.

Carletta, J. (1996) Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics 22, 249_254

Duarte, I. (2000) *Língua Portuguesa. Instrumentos de Análise*. Lisbon: Universidade Aberta.

Eklund, R. (2004) Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues. PhD thesis, University of Linköpink.

Jones, D., Gibson, E., Shen, W., Granoien, N., Herzog, M., Reynolds, D., and Weinstein, C. (2005a). Measuring human readability of machine generated text: three case studies in speech recognition and machine translation. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05), volume 5, pp. 1009–1012.

Landauer, T. K., Dumais, S. T. Dumais (1997) A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review* 104 (2), 211–240.

Liu, Y., Shriberg, E., Stolcke, A., Dustin, H., Ostendorf, M. and Harper, M. (2006) Enriching Speech Recognition with Automatic Detection of Sentence Boundaries and Disfluencies. In *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n. 5, pp. 1526-1540.

Meinedo, H. (2008) *Audio Pre-processing and Speech Recognition for Broadcast News*. PhD. Thesis, Instituto Superior ecnico.

Meinedo, H., Caseiro, D., Neto, J. P., and Trancoso, I. (2003) AUDIMUS.media: A broadcast news speech recognition system for the European Portuguese language. In *PROPOR'2003*, volume 2721 of LNCS, pages 9–17. Springer.

Meinedo, H., Abad, A., Pellegrini, T., Trancoso, I. & Neto, J. (2010) The L2F Broadcast News Speech Recognition System. In *Fala2010*, Vigo, Spain.

Moniz, H. (2006) *Contributo para a caracterização dos fenómenos de (dis)fluência em português europeu.* Master thesis, Faculty of Letters, University of Lisbon.

Moniz, H., Batista, F., Meinedo, H., Abad, A., Trancoso, I., Mata A. I., Mamede, N. (2010) Prosodically-based automatic segmentation and punctuation, In Speech Prosody 2010, ISCA, Chicago, USA, May 2010.

Moniz, H., Batista, F., Trancoso, I. and Mata, A. I. (2011) Analysis of interrogatives in different domain. In *Towards Autonomous, Adaptative, and Context-Aware Multimodal Interfaces: Theoretical and Practical Issues*. Third Cost 2102 International Training School, Springer Berlin/Heidelberg, Lecture Notes in Computer Science, pp. 136-148, Caserta, Italy.

Moniz, H. (2013) *Processing disfluencies in European Portuguese.* PhD thesis, University of Lisbon.

Neto, J., Meinedo, H., Amaral, R., Trancoso, I. (2003) The development of an automatic system for selective dissemination of multimedia information. In: *International Workshop on Content-Based Multimedia Indexing*.

Neto, J., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. 2008. Broadcast news subtitling system in Portuguese. In *Proceeding of ICASSP'08*, pages 1561–1564.

Ostendorf, M., Favre, B., Grishman, R., Hakkani-Tür, D., Harper, M., Hillard, D., Hirschberg, J., Ji, H., Kahn, J., Liu, Y., Makey, S., Matusov, E., Ney, H., Rosenberg, A., Shriberg, E., Wang, W. and Wooters, C. (2008) Speech Segmentation and Spoken Document Processing. In *IEEE Signal Processing Magazine*, pp. 59-69.

Ribeiro, R. and de Matos, D. (2013). Multi-source, Multilingual Information Extraction and Summarization. In *Improving Speech-to-Text Summarization by Using Additional Information Sources*, Springer.

Ribeiro, R., Oliveira, L., and Trancoso, I. (2003). Using Morphossyntactic Information in TTS Systems: comparing strategies for European Portuguese. In *Computational Processing of the Portuguese Language: 6th International Workshop*, *PROPOR'2003*, pages 26–27. Springer.

Shriberg, E. (1994) *Preliminaries to a Theory of Speech Disfluencies*. PhD thesis, University of California.