The LECTRA Corpus – Classroom Lecture Transcriptions in European Portuguese

Isabel Trancoso[1], Rui Martins[1], Helena Moniz[2], Ana Isabel Mata[3], M. Céu Viana[3]
(1) L²F INESC-ID/IST, (2) L²F INESC-ID/CLUL, (3) CLUL

(1,2) INESC-ID Lisbon, Spoken Language Systems Lab, R. Alves Redol, 9, 1000-029 Lisboa, Portugal. {imt, rui.martins, helenam}@l2f.inesc-id.pt
(3) CLUL-Centro de Linguística da Universidade de Lisboa, Complexo Interdisciplinar, Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal. aim@fl.ul.pt, mcv@clul.ul.pt

Abstract

This paper describes the corpus of university lectures that has been recorded in European Portuguese, and some of the recognition experiments we have done with it. The highly specific topic domain and the spontaneous speech nature of the lectures are two of the most challenging problems. Lexical and language model adaptation proved difficult given the scarcity of domain material in Portuguese, but improvements can be achieved with unsupervised acoustic model adaptation. From the point of view of the study of spontaneous speech characteristics, namely disfluencies, the LECTRA corpus has also proved a very valuable resource.

1. Introduction

This paper aims at a detailed description of the corpus collected within the project LECTRA[1]. The goal of LECTRA is to do lecture transcriptions which can be used not only for the production of multimedia lecture contents for e-learning applications, but also for enabling hearing impaired students to have access to recorded lectures.
Lecture transcription can be very challenging, mainly due to the fact that we are dealing with a very specific domain and with spontaneous speech. This topic has been the target of much bigger research projects such as Japanese project described in Furui *et al.* (2001), the European project CHIL (Lamel *et al.*, 2005), and the American iCampus Spoken Lecture Processing project (Glass *et al.*, 2007). It is also the goal of the Liberated Learning Consortium[2], which fosters the application of speech recognition technology for enhancing accessibility for students with disabilities in the university classroom. In some of these projects, the concept of lecture is different. Many of our classroom lectures are 60-minutes long are quite informal. This contrasts with the 20-minute seminar used by Lamel *et al.* (2005), where a more formal style was used.

2. Data collection

The lecture recording effort in the Technical University of Lisbon (IST) is relatively recent and has been done on an experimental basis, restricted to one or two courses per semester. The corpus collected within LECTRA project includes, so far, five 1-semester courses (Table 1). On purpose, we selected very different courses, in order to analyse the influence of several factors.
The PMC course served as the basis for the related Virtual Curricula project that developed a lecture browser application. The course was taught using slides. It is characterized by a very high frequency of computer jargon in English (2.1%), and spelt or partially spelt acronyms (*e.g.*, http).

[1] http://www.l2f.inesc-id.pt/imt/lectra/
[2] www.liberatelearning.com

Acronym	Name	# Lect.	Dur. (h)
PMC	Production of Multimedia Contents	17	16.6
ETI	Economic Theory I	17	15.1
LA	Linear Algebra	27	23.2
IICT	Introduction to Inf. and Comm. Techniques	5	1.8
OOP	Object Oriented Programming	15	18

Table 1: The LECTRA corpus.

The ETI course was one of the courses attended by a student with hearing disabilities. The course was also taught using slides.
The LA course was taught using a white board. The most differentiating factor in this course was obviously the presence of many mathematical variables.
The IICT course was taught in another university. It differs from the other courses in the fact that it was targeted for an internet audience, having been recorded by directly looking at a camera, in a quiet office environment. Given the topic of the course, it also includes much computer jargon in English.
The OOP course is being collected this semester. The lectures are close to 90 minutes each. The course is mostly taught using a white board, although slides with program code are also frequently used. This course represents our first attempt at on-line recognition, *i.e.*, besides recording the video course, the audio is also in parallel fed into our recognizer.
The recording conditions of our four IST courses are similar. The lapel microphone, used almost everywhere, has obvious advantages in terms of non-intrusiveness, but the high frequency of head turning causes audible intensity fluctuations. The use of the head-mounted microphone in the last 11PMC lectures clearly improved this problem. However, this microphone was used with an automatic gain control, causing saturation in 11% of the recordings, due to the increase of the recording sound

level during the students' questions, in the segments after them.
Except for the IICT course, all courses show a relatively high interactivity with the students. This is shown, for instance, by the high frequency of tag question, such as *"não é?"* (isn't it?). In ETI, this particular tag question was used almost once per minute.The variety of possible phonetic realizations in casual speech, coupled with the virtual non-representativeness of such examples in the written corpus, makes them very difficult to be recognized. The fact that we cannot yet correctly place question marks increases the problem, since the negative may be wrongly connected with the next sentence. As expected, no tag questions were found in the non-interactive IICT course.
Discourse markers are also very frequent, the most typical one being *"portanto"* ('so' – 104 instances in 36 minutes, in ETI), pronounced mostly in very reduced forms.
Disfluencies, however, are the most challenging aspect of this type of corpus. The disfluency rate - one in every 10 words, for the PCM course (Trancoso *et al.*, 2006) – is higher than the one cited in Shriberg (2005).

3. Data transcription

The corpus has been manually transliterated using the *Transcriber* tool[3] and the automatically produced text. At this stage, speech is segmented into chunks delimited by silent pauses. Very large chunks may however occur in which no clear silent pause is visible. In such cases, extra boundaries are inserted in order to avoid stretches longer than 10 seconds, preferably at the end of a sense unit and/or at locations where a prosodic break is perceived.
At present, the number of manually transcribed lectures varies very much from course to course. Some courses, such as IITC, due to their restricted size, have been totally transcribed, whereas for others, the orthographic transcription is not yet completed.
Given the richness of this corpus for the study of a variety of speech phenomena, disfluencies in particular, a subset of the corpus, including around 2 hours from each course, was fully annotated at different levels. The multilayer manual annotation was done using the *Wavesurfer* tool[4] and additional information obtained automatically is provided in compatible format.

3.1 Orthographic transcription

An enriched orthographic transcription (Snover et al., 2004) is crucial, not only to make the recognition output intelligible for readers (e.g. hearing impaired students), but also to facilitate the study and testing of automatic capitalization and punctuation methods. The first annotation tier contains, thus, the full transcription of what has been said by the teacher such as described above, enriched with punctuation marks. The use of punctuation was restricted, however, to full stop, comma, and question mark, in order to clearly distinguish declarative and interrogative sentence-like units, avoiding more complex and subjective punctuation markers.
Disfluent sequences that may eventually be suppressed are delimited by angular brackets and their edges aligned with the speech signal. Those sequences are transposed to the disfluency tier, where more detailed analysis is provided.
Segmentation marks were also inserted for regions in the audio file that were not further analysed. Those are delimited by square brackets and may correspond to changes in background noise, signal saturation, speech overlap, etc. Special attention has been given to the marking of students' comments and clarification requests, which are inaudible, given the recording conditions, but crucial to the written text readability.
Other symbols at the left and right edge of the words were also added, to account for different types of phenomena and /or facilitate subsequent searches. Table 2 shows the complete list of used symbols.

Symbols	Context of Use
< >	Auto-corrected sequences
[ ]	Non analyzable speech sequence delimiters
^	Proper names whose first letter must be capitalized,
~	Right edge of the word – irregular pronunciation; Left edge - spelled sigla or mathematical expression/variable
@	Siglae an Acronyms pronounced as regular words
+	Word contractions or syncopated forms
§	Morfosyntactic irregular forms.
%	Filled pauses
-	Word fragment
=	Excessive segmental prolongations

Table 2: Symbols used in the orthographic tier.

In previous studies for E.P. (Mata, 1999; Moniz, 2006), including our preliminary work on university lectures (Trancoso *et al.*, 2006) only three basic forms have been found for filled pauses: an elongated central vowel only, a central vowel followed by a nasal murmur and a nasal murmur only, which have been orthographically coded as "aa", "aam" and "mm", respectively. More recent studies, however, (Moniz, *et al*, 2007) refer a strong speaker dependent variation in what the form and contextual distribution of filled pauses is concerned. While some speakers appear to use just one form in all contexts (a central vowel only), others may used the three forms and the quality of the vowel may also change according in function of the height of the previous word last vowel. In order to get a better insight into this type of variation, and also to distinguish between long and brief forms which appear to accomplish distinct functions, those were coded using the SAMPA alphabet for European Portuguese, preceded by "%".

3.2 Disfluency annotation

The annotation of disfluencies is provided in a separate tier (annotation file), closely following Shriberg (1994) and basically using the same set of labels, as shown in table 3.This annotation scheme, is based on Levelt's (1983) model and has been successfully used to train methods for the identification and automatic removal of disfluencies in order to produce clean readable texts

[3] http://trans.sourceforge.net/
[4] http://www.speech.kth.se/wavesurfer

(Shriberg, 2005 and references therein). It appears to be also the most adequate from a point of view of linguistic research as, in spite of some divergences, it is widely accepted that disfluencies have an internal structure and, three different regions at least, need to be consider in their analysis: (i) the *reparandum,* containing the linguistic materials to be repaired; (ii) the *interregnum* or temporal region of variable length which may contain filled or unfilled pauses, as well as editing terms; and (iii) the *repair* itself. The *reparandum* is right delimited by an *interruption point*, marking the moment in time in which an interruption is visible in the surface form.

Labels	Description
< >	Auto-corrected sequences delimiters
.	Interruption point
f	Filled pauses
lm	Segmental prolongations
r	Repetitions
s	Substitutions
d	Deletions
i	Insertions
e	Editing terms/expressions
p	Discourse markers (as fillers)
-	Word fragment
~	Mispronunciations

Table 3: Labels used in the disfluency tier.

In our transcription, the left and right boundaries of each of these regions are aligned with the speech signal. An additional label (lm) was added however to Shriberg's tag set, as strong evidence was found in our previous work (Moniz *et al.*, 2007) that they should be treated at pair with filled pauses. In fact, although further work is needed in order to get a better insight into the specific behaviour of filled pauses and prolongations, results obtain so far support the view that both phenomena can be regarded as manifestations of planning effort at different levels of the prosodic structure and, therefore, considered to be in complementary distribution. Independent evidence for the treatment of prolongations at pair with other classes of disfluencies was also presented in Eklund (2004), whose results suggest language specific effects on their production and contextual distribution. Following a suggestion of the latter author, disfluent items are indexed, as shown in the following example.

ort	<tem	um	número	tem	um	número,	não.>	tem	um	elemento.		
	has	*a*	*number.*	*has*	*a*	*number,*	*no.*	*has*	*an*	*element*		
disf	< r1	r2	s1	• r1	r2	s1	e1	r1	r2	s1	>	

Such a solution appears to be less pruned to errors than the complex bracketing used by Shriberg in order to account for the nested structure of long disfluency sequences. Unlike Eklund, however, all items are indexed for a more direct access to eventual changes in word order and to the different strategies that may be used by speakers.

3.3 Syntactic annotation

The third level of manual annotation aimed both at providing basic syntactic information and a segmentation of the speech string into Sentence-like Units (SUs), closely following LDC guidelines.[5] As a representation of the later in standard writing format would constitute an unnecessary reduplication, two basic labels were used, instead: SU for sentence-level breaks and SI for sentence internal one. Those are followed by an underscore, after which the type of clause is indicated (e.g. SI_RR for restrictive clauses, SU_CC and SI_CC for coordination of main clauses and clauses with semantic dependency, respectively, etc).

3.4 Automatic annotation

Once the orthography is ready and the non-analysable regions demarcated, a WFST-aligner which can cope with alternative pronunciation rules (Trancoso *et al*, 2003, and references therein) is used to generate two additional annotation files: one with words boundaries and the other one with phone boundaries. Phones are given in SAMPA for European Portuguese.

Additional morphosyntactic information is currently being added, which is built upon the following tools: (i) Palavroso/MARV, a POS tagger coupled with a disambiguator; (ii) RUDRICO, an improved version of PASMO (Pardal & Mamede, 2004) which. When applied twice splits and concatenates tokens; (iii) XIP (Ait-Mokhtar *et al*., 2001) that returns the input organized in chunks, connected by dependency relations.

4. Related text materials

The fact that the lectures are taught in European Portuguese adds further challenges. A common denominator of many technical courses in Portugal is the adoption of textbooks in English, namely for Master level courses. Hence, it is difficult to find text materials to adapt the recognizer to the course domain.

For the PMC course, the textbook was in English. Our material in Portuguese was thus very restricted: slides (25k words), exam questions (2k words) and student reports (23k words).

Slides are typically characterized by specific grammatical constructions which clearly differentiates them from other textual sources. By analysing a small set of PMC material, we obtained a much smaller percentage of verbs in slides *vs.* reports (9.1% *vs.* 17.0%) and a much higher percentage of nouns (42.2%*vs*. 27.1%).

For the ETI course, we had a textbook (452k words), we discarded the slides. For the LA course, the internet material amounted to around 80k words. For the IICT course, we have introductory slides (4k words) and web notes (61k words). For the OOP course, we have mostly notes from the still growing wiki page of the course.

5. Baseline Recognizer

Our baseline large vocabulary recognizer (Trancoso *et al.*, 2003) was trained for Broadcast News (BN). It has a vocabulary of around 57k words, a perplexity (PP) of 139.5, and an out-of-vocabulary (OOV) word rate close to 1.4%. For BN, this hybrid recognizer achieves a word error rate (WER) of 31.2% for all conditions, 18.9% for

[5] http://www.ldc.upenn.edu/Projects/MDE

read speech in studio, and 35.2% for spontaneous speech. Applying this recognizer to our small test set (one single lecture per course), without any type of domain adaptation, obviously yields very bad results, as shown in Table 2.

An analysis of the main types of recognition errors revealed several sources which were common to the BN domain: disfluências, severe vowel reduction, OOVs, large variability of inflected forms, and inconsistent spelling. Disfluencies play a major role, being responsible for the frequent error bursts (only 19.2% of the errors in PMC occurred in isolation).

Course	WER (%)	PP	OOV (%)
PCM	63.6	292.8	3.4
ETI	56.4	175.0	1.6
LA	67.0	330.9	5.2
IICT	43.2	286.5	4.2
OOP	79.6	253.9	3.1

Table 2: Baseline results.

6. Domain Adaption

Our first efforts at domain adaptation were described by Trancoso *et al.* (2006) for the PMC and ETI courses. After lexical and language model adaptation, the results are still very poor (relative 7.7% and 3.7% reduction in WER, respectively). After combining this first adaptation stage with supervised acoustic model adaptation, the relative reduction is much more significant (29.6% and 20.7%, respectively), but the resulting WER is still very high.

This motivated the study of unsupervised acoustic model adaptation. We first tested off-line adaptation of single lectures. Even for the very short lectures of the IICT course, this already resulted in a relative WER reduction of 6.5%.

With the target of on-line recognition, we next tried using in a new lecture, the acoustic models adapted with the material from previous lectures recognized with a high confidence level. For the last IITC lecture, for instance, the relative WER reduction was 10.6%. However, using more than 35 minutes of previously recognized material for this course did not result in WER improvements.

7. On-line versus off-line Recognition

As explained, the OOP course was our first on-line recognition experiment with lectures. Most of the work done in that direction was based on the automatic captioning work we are currently doing for the national TV station, and on the development of our dictation system. During the first lectures, we recorded the audio material for unsupervised acoustic domain adaptation.

On-line recognition poses latency problems. Although we have not yet conducted on-line test with courses based on slides, we believe that the very frequent pointing at the board may be even more problematic, given the 4-5 words latency.

8. Conclusion

This paper described our small corpus in European Portuguese, and the problems it raises for automatic speech recognition systems. The fact that a significant percentage of the recognition errors occurs for function words (45.1%) led us believe that the current performance, although far from ideal, may be good enough for information retrieval purposes, enabling keyword search and question answering the lecture browser application.

The LECTRA corpus is also being currently used to study the application of punctuation and capitalization methods (Batista *et al.*, 2007). We believe that producing a surface rich transcription is essential to make the recognition output intelligible for hearing impaired students.

9. Acknowledgements

The authors would like to thank João Neto, Hugo Meinedo, Ciro Martins, and Joaquim Jorge, for many helpful discussions. This work was partially funded by FCT project POSC/PLP/58697/2004 and by PRIME National Project TECNOVOZ number 03/165. INESC-ID Lisboa had support from the POSI Program of the Quadro Comunitário de Apoio III.

10. References

Ait-Mokhtar, S., Canhod, J., Roux, C. (2004). Multi-input Dependency parser. In *Proceedings of the 7th Workshop on Parsing Technologies*. Beijing, China.

Batista, F., CAseiro, D., Mamede, N., Trancoso, I. (2007). Recovering Punctuation Marks for Automatic Speech Recognition. In *Proceedings of Interspeech'2007*, Antwerp, Belgium.

Eklund, R. (2004). Disfluency in Swedish Human-Human and Human-Machine Travel Booking Dialogues. Ph.D theses. Institute of Technology, Linköping University.

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., Tamura, S. (2001). Ubiquitous Speech Processing. In *Proceedings of ICASSP'2001*, Salt Lake City, USA.

Glass, J. *et al.* (2007). Recent Progress in the MIT Spoken Lecture Processing Project. In *Proceedings of Interspeech'2007*, Antwerp, Belgium.

Lamel, L., Adda, G., Bilinski, E., Gauvain, J. (2005). Transcribing Lectures and Seminars. In *Proceedings of Interspeech'2005*, Lisbon, Portugal.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. Cambridge University Press.

Mata, A. I. (1999). *Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implcações Didácticas.*

Medeiros, J. (1995). Análise Morfológica e Correcção Ortográfica do Português. M.S. Dissertation. Instituto Superior Técnico, Universidade Técnica de Lisboa.

Moniz, H. (2006). *Contributo para a Caracterização dos Mecanismos de (Dis)Fluência no Português Europeu*. Ma. Faculdade de Letras da Universidade de Lisboa..

Moniz, H., Mata, A. I., Viana, M. C. (2007). On Filled Pauses and Prolongations in European Portuguese. In *Proceedings of Interspeech'2007.* Antwerp, Belgium.

Pardal, J., Mamede, N. (2004). Terms Spotting with Linguistics and Statistics. In *Proceedings of the Herramientas y Recursos Linguísticos para el Espanol y el Português Workshop*. Pp. 298-304.

Ribeiro, R., Mamede, N., Trancoso, I. (2004). Using Morphosyntactic Information in TTS Systems: Comparing Strategies for European Portuguese. In *PROPOR'2003 – 6th Workshop on Computational Processing of the Portuguese Language*.

Springer-Verlag, Heidelberg, pp. 143-150.
Shriberg, E. (1994). Preliminaries to a Theory of Speech Disfluency. Ph.D theses. University of California.
Snoven, M., Schwartz, R., Dorr, B., Makhoul, J. (2004). RT-S: Surface Rich Transcription Scoring Methodology, and Initial Results. In *Proceedings of the Rich Transcription 2004 Workshop*. Montreal, Canada.
Trancoso, I., Caseiro, D., Viana, M. C., Silva, F., Mascarenhas, I (2003). Pronunciation Modeling Using Finite State Transducers. In *ICPhS'2003 – 15th International Congress o f Phonetic Sciences*. Barcelona.
Trancoso, I., Neto, J., Meinedo, H., Amaral, R. (2003). Evaluation of na Alert System for Selective Dissemination of Broadcast News. In *Proceedings Eurospeech'2003*, Geneva, Switzerland.
Trancosos, I., Nunes, L., Neves, L., Viana, M., Moniz, H., Caseiro, D., Mata, A. (2006). Recognition of Classroom Lectures in European Portuguese. In *Proceedings of Interspeech'2006*, Pittsburgh, USA.