Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4293–4301
Marseille, 11–16 May 2020

A Multimodal Educational Corpus of Oral Courses: Annotation, Analysis and Case Study

Salima Mdhaffar[1,2], Yannick Estève[2], Antoine Laurent[1], Nicolas Hernandez[3], Richard Dufour[2], Delphine Charlet[4], Géraldine Damnati[4], Solen Quiniou[3], Nathalie Camelin[2]
[1] LIUM - University of Le Mans, France
[2] LIA - University of Avignon, France
[3] LS2N - University of Nantes, France
[4] Orange, France
{firstname.lastname}@univ-lemans.fr,univ-avignon.fr,univ-nantes.fr,orange.com

Abstract

This paper presents a French multimodal educational dataset of oral courses. This corpus is part of the PASTEL (Performing Automated Speech Transcription for Enhancing Learning) project aiming to explore the potential of synchronous speech transcription and application in specific teaching situations (Bettenfeld et al., 2018; Bettenfeld et al., 2019). It includes 10 hours of different lectures, manually transcribed and segmented. The main interest of this corpus lies in its multimodal aspect: in addition to speech, the courses were filmed and the written presentation supports (slides) are made available. The dataset may then serve researches in multiple fields, from speech and language to image and video processing. The dataset will be freely available to the research community. In this paper, we first describe in details the annotation protocol, including a detailed analysis of the manually labeled data. Then, we propose some possible use cases of the corpus with baseline results. The use cases concern scientific fields from both speech and text processing, with language model adaptation, thematic segmentation and transcription to slide alignment.

Keywords: Multimodal corpus, Educational context, Thematic segmentation, Alignment, Language model adaptation

1. Introduction

With the increasing number of applications handling spontaneous speech, lecture processing has becoming an active field of research. In this particular educational context, a large number of projects have been developed, coming with different datasets. Among them, we can first quote the LECTRA corpus (Trancoso et al., 2006; Trancoso et al., 2008) dealing with classroom lectures in European Portuguese. This corpus provides the audio of lectures and their manual transcription. In addition to the oral modality, the “Spontaneous Speech Corpus and Processing Technology”(Furui et al., 2000; Furui et al., 2001) and the CHIL projects (Lamel et al., 2005) include, in addition to the audio and the transcription, the video recording of lectures. Finally, the LMELectures corpus (Riedhammer et al., 2013) is the most complete one with various modalities (audio, video and text), including the annotation of speech transcription, a segmentation in speaker turn, as well as keywords definition.

We propose in this paper an original French speech educational corpus that includes 3 basic modalities: the oral lecture (audio), video recording and the presentation supports (text). To these modalities are included additional manual annotations: manual transcriptions of lectures, in-domain words extraction and annotation, and alignment of presentation supports (slides) and oral presentation during the lectures. To our knowledge, there is no existing corpus that integrates such a variety of annotations.

This paper aims at giving a detailed description of the corpus collected within the PASTEL project[1] (Performing Automated Speech Transcription for Enhancing Learning). We expect that this corpus, dedicated to the educational field and freely released for research and industrial communities, will allow new advances in this particular context by proving a general framework to develop, experiment and evaluate systems on various applications and domains.

The paper is organised along the following lines. Section 2. presents the sources of the collected data. Section 3. details the annotation guidelines. The annotation statistics and analysis are respectively presented in Section 4. and 5. The case study is described in Section 6., before concluding and giving some perspectives in Section 7.

2. PASTEL Corpus Datasource

The PASTEL corpus consists in a collection of courses in various computer science fields (automatic language processing, introduction to computer science, etc.) in the first year of a computer science degree at the University of Nantes. Two sources were used to collect data from the PASTEL corpus: the CominOpenCourseware (COCo) and the Canal-U platforms. Courses whose source is COCo platform were filmed and posted on a web platform[2] following the COCo project which ran from 2013 until 2016. The main goal of this project was to mobilize video annotations in pedagogical and research contexts, and to promote open educational resources and open licenses (Aubert et al., 2014; Aubert and Jaeger, 2014; Mougard et al., 2015; Canellas et al., 2015). Six courses have been downloaded from the platform of this project. Three other courses have been downloaded from the Canal-U platform[3]. Canal-U is a site containing audiovisual resources of higher education

[1] https://projets-lium.univ-lemans.fr/pastel/

[2] http://www.comin-ocw.org/

[3] https://www.canal-u.tv/

and research, fed mainly by academics and French universities.

Figure 1 shows a screenshot of the COCo platform. Each course contains the course video, course material and an alignment between the video and the slides of the course material. Courses whose original source is Canal-U (3 lectures) do not have the lecture support and therefore no alignment information.

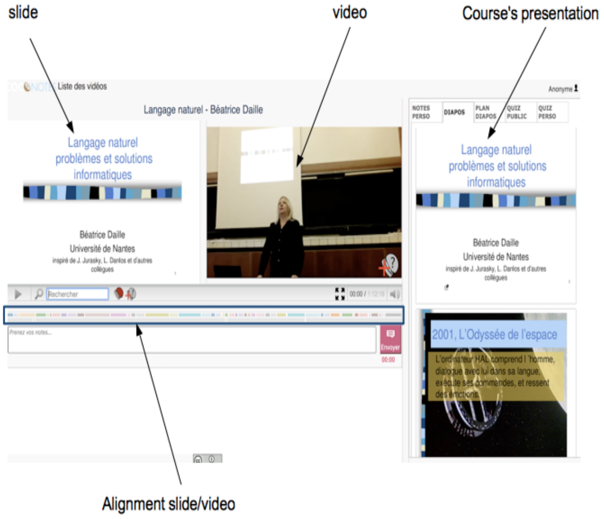

Figure 1: Screenshot of the COCo platform for the lecture *"Natural language processing"*

3. Annotation Protocol

We present in this section the annotation protocol, including manual transcription (Section 3.1.), thematic segmentation (Section 3.2.), and in-domain words annotation (Section 3.3.), established for annotating the PASTEL corpus.

3.1. Manual transcription

This manual transcription was carried out in two passes. The first pass consists in automatically transcribing the course through a generic speech recognition system (*i.e.* not adapted to the targeted courses). The human expert intervenes during the second pass to manually correct the errors made by the automatic transcription. This two-step approach proved to be a faster way than a manual *from scratch* transcription (Bazillon et al., 2008). The conventions used for the evaluation of transcription campaigns (Gravier et al., 2004) served as a guide for manually transcribing registered lectures.

The speech recognition system is based on the Kaldi toolkit (Povey et al., 2011). Acoustic models were trained on about 300 hours of speech from French broadcast news with manual transcriptions, and are based on a chain-TDNN approach (Povey et al., 2016). The generic n-gram language models were trained on these manual transcriptions of speech, but also on newspaper articles, for a total of 1.6 billions of words. The vocabulary of the generic language model contains around 160k words. More details about language models can be found in (Rousseau et al., 2014).

The Transcriber tool was used (Barras et al., 2001) to perform the manual transcription. Transcriber is an optimised software for the transcription and the annotation of large corpora.

3.2. Thematic segmentation

In order to guide thematic segmentation inside lectures, we must answer the following question: "What is a thematic segment?" in a course which is supposed to be monothematic (*i.e.* has a main subject, *Introduction to computer science*).

After a long study, we assumed that a thematic boundary can only be in the vicinity of a slide change during the course. Therefore, for each change of slide, a human expert annotated:

1. If there is a topic shift.
2. The exact moment of the topic shift defined as being positioned between two words.
3. The granularity of the topic shift (1 or 2) or if the segment type is an interruption.

Granularity 1 is used to mark the beginning of a new notion in the course while staying in the same global topic. Granularity 2 is used when there is a more general sub-thematic change that allows to stop learning at that time and continue later learning other notions.

Out of these topic granularities, interruptions, which correspond to moments of public management or technical problems (e.g. video-projector trouble shouting), have been annotated.

The annotation was carried out by two students in Master's degrees in linguistics using the ELAN software (Auer et al., 2010), a linguistic annotation tool designed for creating text annotations for audio and video files.

3.3. In-domain words annotation

In-domain words correspond to the linguistic expressions which refer to concepts, objects or entities being essential for the understanding of the current slide or a given transcription. We have included all the scientific and technical terms as well as acronyms and expressions allowing us to go further in the course topic. In-domain words were manually extracted from both manual transcriptions and presentation slides. This annotation was made only for courses for which slides were provided.

4. Corpus Statistics

Table 1 presents some statistics about the lectures[4]. The second, third, and fourth columns of the table represent the numbers of "granularity 1" , "granularity 2" and "interruption" segments, respectively. Column 5 and 6 present respectively the number of annotated in-domain words for both transcriptions and slides. Column 7 presents the number of slides for each lecture. The last column contains the duration of each lecture. The number of speakers in

[4] Lecture's names: *(1) Introduction to computer science*, *(2) Introduction to algorithms*, *(3) Functions*, *(4) Social networks and graphs*, *(5) Distributed algorithms*, *(6) Natural language processing*, *(7) Republic Architecture*, *(8) Traditional methods*, *(9) Imagery*

this corpus is 7. Lectures (1), (2) and (3) have the same speaker. As said in previous section, note that 3 lectures ((7), (8) and (9)) were made without slides. Lectures from (1) to (6) are recorded from Coco and lectures from (7) to (9) are recorded from Canal-U.

Table 1: Corpus statistics: Duration, number of granularity 1 units (G1), granularity 2 units (G2), interruptions (I), in-domain in transcripts (Kw_t), in-domain in slides (Kw_s), number of slides (#S) and duration (Dur.).

Lec.	**G1**	**G2**	**I**	Kw_t	Kw_s	#S	**Dur.**
(1)	31	2	2	47	54	75	1h 04m
(2)	38	10	3	25	35	62	1h 17m
(3)	35	3	3	109	78	137	1h 14m
(4)	42	7	7	54	84	64	1h 05m
(5)	72	5	3	232	146	73	1h 16m
(6)	52	5	5	120	100	55	1h 09m
(7)	49	7	0	-	-	-	1h 21m
(8)	12	7	1	-	-	-	0h 41m
(9)	57	0	1	-	-	-	1h 08m
Total	388	46	25	587	497	466	10h25m

5. Annotation Analysis

In the previous sections, we presented the different sources used to collect the corpus and the protocol followed to annotate the data. In this section, we focus on describing qualitatively and quantitatively the annotated data, by analysing first the thematic segmentation annotation (Section 5.1.), and second the annotated in-domain words (Section 5.2.).

5.1. Analysis of thematic segmentation

As presented in Section 3.2., a thematic segment consists in speech related to one or more slides. The duration of a segment can range from a few seconds to tens of minutes. Tables 2 and 3 present the segment duration statistics, globally and for each individual course, for Granularity 1 and 2 respectively. The second column represents the number of segments. The third column represents the average duration of the segments. Columns 5 and 6 respectively represent the minimum and the maximum duration among the segments of each course.

Table 2: Number and duration (average, minimum and maximum (ms)) of segments of Granularity 1 per lecture.

Lecture	**G. 1**	**av-dur**	**min-dur**	**max-dur**
(1)	31	123.0	16.3	307.8
(2)	38	106.6	18.7	248.4
(3)	35	124.3	42.2	393.8
(4)	43	85.3	11.6	475.6
(5)	72	53.8	6.4	204.4
(6)	52	80.3	5.2	215.2
(7)	49	92.3	14.4	317.6
(8)	12	187.2	17.4	724.5
(9)	57	63.0	4.0	224.1

Statistics in Tables 3 and 2 highlight the large disparity in the size of the segments between the different courses but also within the same course.

Table 3: Number and duration (average, minimum and maximum (ms)) of segments of Granularity 2 per lecture.

Lecture	**G. 2**	**av-dur**	**min-dur**	**max-dur**
(1)	2	455.6	455.6	455.6
(2)	10	476.1	129.9	1041.5
(3)	3	1036.5	584.6	1672.5
(4)	5	960.8	285.6	1871.1
(5)	5	1114.7	466.4	1824.4
(6)	5	960.8	285.6	1871.1
(7)	7	696.3	350.5	1179.6
(8)	7	340.6	45.4	874.3
(9)	0	-	-	-

5.2. Analysis of manual in-domain words

In-domain words constitute an important part in the PASTEL corpus. In this context, we propose to study in this section:

- *Occurrence of in-domain words*. The occurrence of a in-domain words refers to the number of its apparition in the corpus.
- *Distribution of in-domain words in slides*. The distribution of a in-domain words in a corpus is the set of locations where this expression appears: do they appear uniformly throughout the corpus or are they rather localised in a few slides ?

5.2.1. Occurrence of in-domain words

Figure 2 presents a simple calculation of the occurrences of in-domain words annotated from the course presentation in both manual transcription and course presentation for the lecture *"introduction to computer science"*. For sake of comparison, Figure 3 presents the same calculation but for in-domain words annotated from manual transcriptions. The number of occurrences of in-domain words in manual transcriptions is represented by red bars. The number of occurrences of in-domain words in the course material is represented by blue bars.

Figures 2 and 3 show that the key-expressions are different in terms of occurrence and ubiquity. It is observed that the number of occurrences of in-domain words is not similar in the slides and in the teacher's speech. We also observe that many in-domain words appear only once in the corpus. Note that the same behaviour has been observed for all lectures in the corpus.

5.2.2. Distribution of in-domain words in slides

As we have announced in Section 3.2., a slide is considered in the context of our data as an atomic textual unit, on the scale of it the corpus has been segmented. According to this, the internal structure of the slides in terms of word distribution is important. Figure 4 presents the distribution of in-domain words for one lecture. The horizontal axis present the in-domain words and the vertical axis present the slides (slide 1, slide 2, ..., slide n) where n is the number of slides in a lecture (column 7 in Table 1). Figure 4 shows that our data suffer from a lack of repetitions of the in-domain words in successive slides. Note that we observed the same behaviour for all lectures. This lack

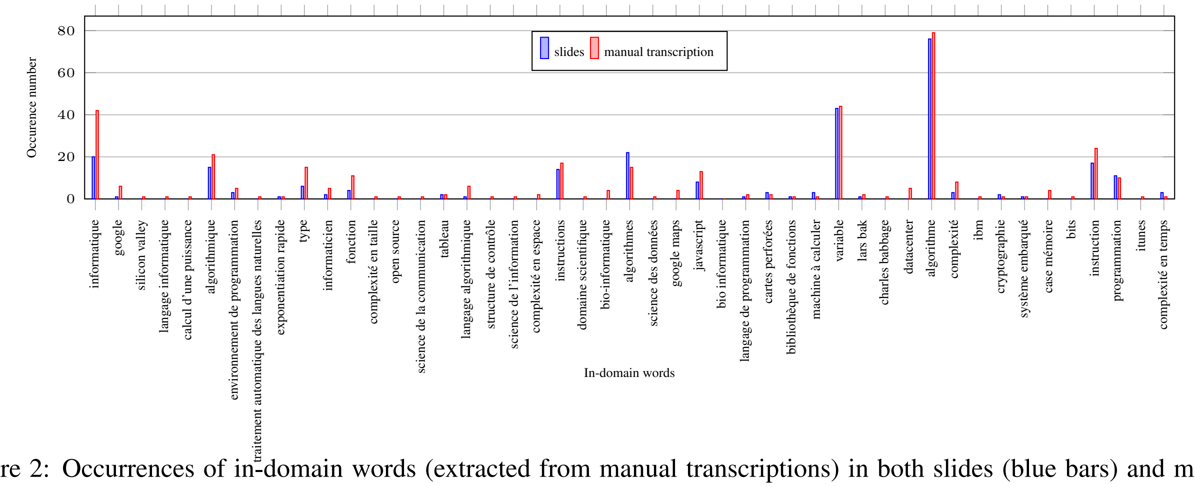

Figure 2: Occurrences of in-domain words (extracted from manual transcriptions) in both slides (blue bars) and manual transcriptions (red bars)

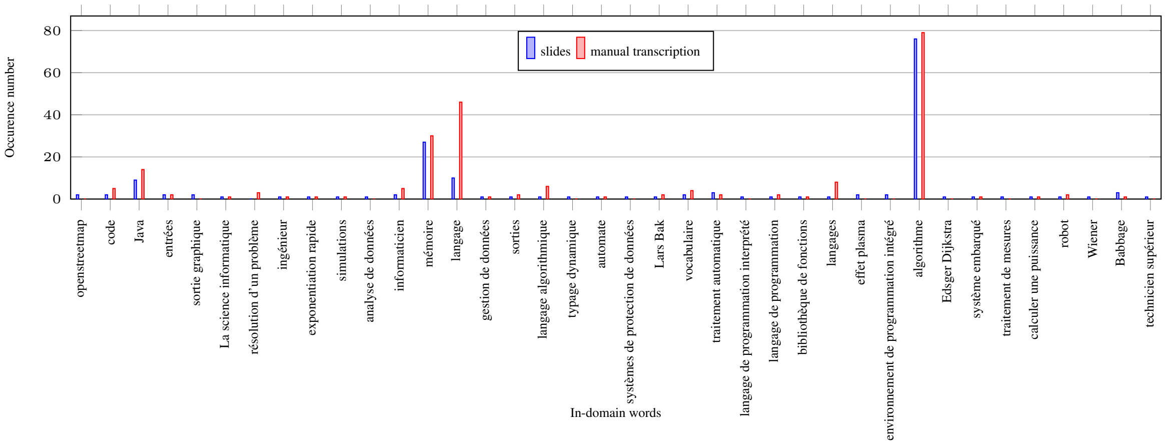

Figure 3: Occurrences of in-domain words (extracted from slides) in both slides (blue bars) and manual transcription (red bars)

of repetition can have a negative impact on the thematic segmentation task. The use of other useful information besides the repetition of words for the segmentation task then becomes a priority that we focus on in Section 6.2.

6. Case Studies

To illustrate the usefulness of this corpus and to provide first baseline results, we have performed some experiments on different language processing task: adaptation of speech recognition systems (Section 6.1.), thematic segmentation of lectures, (Section 6.2.), and temporal alignment of written supports (slides) and oral lectures (Section 6.3.).

6.1. Automatic Speech Recognition (ASR) adaptation

Automatic transcription of lectures can be very challenging, mainly due to the fact that we are dealing with spontaneous speech, but also with a very specific domain. For this last difficulty, language model (LM) adaptation is a commonly used technique to address the mismatch problem between a generic model, trained with a large set of multi-domain textual data, and the targeted data to transcribe, that can be trained on a specific thematic dataset. Our first efforts at domain adaptation were described in details in the paper (Mdhaffar et al., 2019). For automatic speech transcription of lectures, texts of presentation slides are expected to be useful for adapting a language model (Yamazaki et al., 2007; Kawahara et al., 2008; Miranda et al., 2013; Akita et al., 2015). Based on these works, we use slides of a presentation to extract relevant data to collect web data. First, we consider that slide titles are essential for giving listeners a quick idea of the content of a course part. Indeed, this is often the main information on which a listener relies to search and to point out in the course. The idea is then to use slide titles as queries to retrieve web documents. Nonetheless, slide titles are sometimes generic. For example, the use of a query with the slide title *"variable"* of the lecture *"introduction to algorithms"* can give many results out of the targeted topic. As a result, we decided to concatenate each title to the general title of the lecture. Queries are then submitted to a web search engine (Google in our experiments) and the returned page are downloaded. We have limited the search to 100 web pages for each query. Textual content of collected web pages has been extracted and normalised. Finally, we adapt LM by linear interpolation between an existing generic LM and a LM trained with web data. Table 4 presents results of LM adaptation.

The goal of LM adaptation is to improve the performance for in-domain words. The annotated in-domain words in the PASTEL corpus gives the possibility to compute errors for in-domain words using information retrieval metrics such as precision, recall and f-measure, or using a more adapted

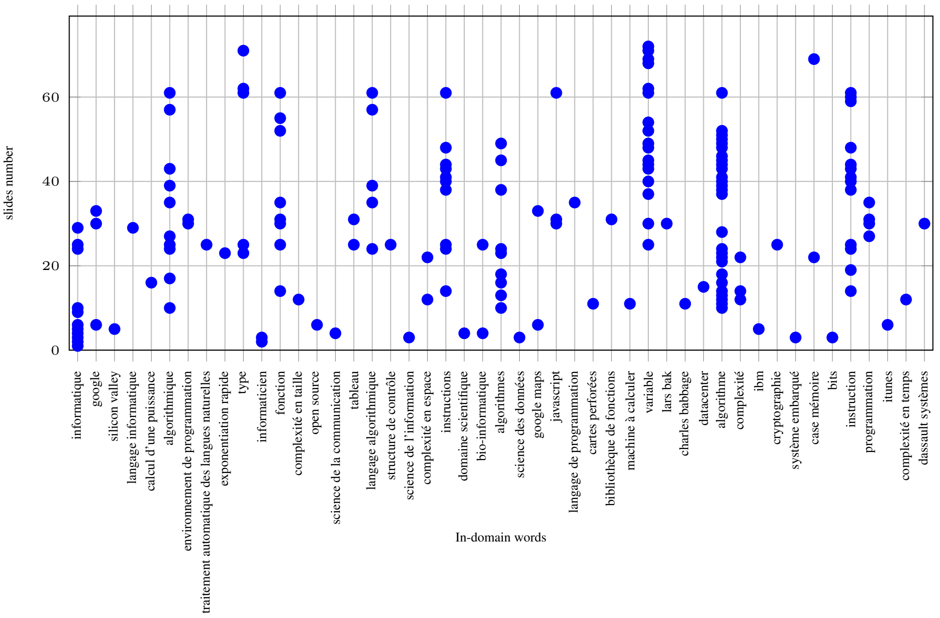

Figure 4: Distribution of annotated in-domain words from the manual transcription of the *"Introduction to computer science"* lecture.

Table 4: (%) WER for LM adaptation on the PASTEL corpus.

	ASR w/o adaptation	**ASR w/ adaptation**
All words (= WER)	19.46	16.42

metric like the $IWER_{Average}$ score proposed in (Mdhaffar et al., 2019). Results on $IWER_{Average}$ are shown in Table 5.

Table 5: (%) $IWER_{Average}$ scores for manual in-domain words annotations from slides and transcripts.

	ASR w/o adaptation	**ASR w/ adaptation**
Manual slide keywords	32.31	14.52
Manual transcript keywords	31	17.30

6.2. Thematic Lecture Segmentation

Recent advances in ASR allow us to imagine new applications for enhancing learning. Lectures can automatically be transcribed in text which, in turn, can be used by learners to read the courses. But, unlike handouts or any written educational resources, transcribed lecture audio can be tedious to browse, making it difficult for learners to retrieve relevant information in the absence of structural information such as topic boundaries. We present in this section our work on the thematic segmentation of oral lectures.

Our thematic segmentation baseline system consists in using the TextTiling algorithm (Hearst, 1997) which is based on analysis of lexical distribution between adjacent blocs. The main reasons of choosing TextTiling is related to its simplicity and that it is an unsupervised algorithm that does not require training data (note that our corpus only contains 388 segments).

In TextTiling, a block is constituted of k sentences, while a sentence is constituted of w words. Similarity is computed using a sliding window between adjacent blocks. The similarity values allow us to draw a lexical cohesion curve. Topic boundaries are detected based on the valley depth of the lexical cohesion curve. Otherwise, those two blocks belong to two different topics.

We have used MACAON to stem the transcription. We compute the cosine similarity between two adjacent blocks using TF-IDF weighting.

Line 1 in Table 6 presents results of our baseline system. The size of w is 10 words. As our algorithm is based on a configurable sliding window, we applied cross-validation to set this value for each course. Results show the difficulty of thematically segmenting teaching courses.

TextTiling algorithm has been developed to segment text. For topic segmentation of TV broadcast news, (Bouchekif et al., 2015) have proposed an algorithm inspired from TextTiling. The similarity is computed between 2 bloks of breath group. A breath group corresponds to the speech spoken by a speaker between two breaths (silent pauses). The valleys are then detected by a recursive mechanism for detecting local minimum. A second pass of the algorithm can be made, considering neither the lexical cohesion but the semantic cohesion between the windows. We have tested this algorithm to segment lectures. Performance is presented in line 2 in Table 6. Results show this algorithm is less effective than TextTiling in this use case.

While manual segmentation is based on slide's change, we decide to compute the performance of segmentation considering each slide change in a lecture as a thematic boundary. As expected, results show that the structure in slides of a lecture gives an important information for the segmentation task. Based on these results, we propose to give a weight to each similarity according to its distance to a slide change. We apply this modification to the best baseline system (TextTiling (line 1)). Performance is presented in line 3 in Table 6. The system performance is improved when taking account of the distance to slide change.

Table 6: Thematic segmentation results on the PASTEL corpus (Precision, recall and F-metric are computed with a tolerance margin of 10 seconds as in (Bouchekif et al., 2017)).

	Precision	Recall	F-metric
TextTiling	27.53	65.66	38.79
(Bouchekif et al., 2015)	40.50	34.40	37.20
Slide's change	48.60	83.39	61.41
SliTextTiling	60.05	85.66	70.60

6.3. Alignment of slides and oral lectures

Another structural information that can help learners to browse the transcription and to point out in a specific part is the alignment of slides and speech transcription. We present in this section our method on the temporal alignment between the oral lecture and the visual text supports (slides). Our proposed method exploits the textual content of slides and speech transcription segments to perform their alignment. Manual and automatic segment transcripts refer to the words contained in this automatic segmentation. The first step of the ASR consists in performing an audio segmentation stage. The generated output corresponds to homogeneous speech segments delimited by silence breaks or speaker shifts (see (Broux et al., 2018)).

The text on the slides, as well as the automatic and manual speech transcriptions, are stemmed using the MACAON[5] NLP tool (Nasr et al., 2011). Stopwords are removed in the slide texts as well as in the speech transcription.

We designed our method to consider: (1) textual similarity between slides and speech transcription segments, and (2) linear ordering of slide and transcription segments. We build a separate module for each of these two analyses and merge them to get the final decision.

6.3.1. Similarity measure between slides and speech transcription

The similarity measure serves as the primary basis to align slides and speech transcription segments. We compute a cosine similarity between slide and speech transcription segments, using TF-IDF weighting. In this work, we build a TF-IDF representation of transcript segments and slides which computes the IDF in regards to their document collection (considering the slides and the transcript segments as two distinct collections).

All the terms used in the slides or in the transcription segments define the vocabulary to compute TF-IDF.

Let $S = \{s_1, s_2 ... s_n\}$ be a set of text slides and $T = \{t_1, t_2 ... t_m\}$ be a set of speech transcription segments, where n is the number of slides and m is the number of speech transcription segments. We define $Sim(t_i, s_j)$ as the cosine similarity between the representation vector of speech transcription segment t_i and the representation vector of slide s_j.

Let be $\Pi = \{\pi_1, \pi_2, ..., \pi_k\}$ the set of all the possible sequences where $\pi_x = [(t_1, s_i), (t_2, s_j), ..., (t_m, s_k)]$ of (slide, speech segments) pairs, with the assumption that each speech segment is aligned with one (and only one) slide, while one slide can be aligned to different segments. For each sequence π_x, we compute the score $L(\pi_x)$, following the formula:

$$L(\pi_x) = \prod_{(t_i, s_j) \in \pi_x} Sim(t_i, s_j) \qquad (1)$$

6.3.2. Slide and speech segment order constraint

The slide and speech segment order constraint is defined to impose a linear structured order on the slides and speech segments.

Let $\alpha = [p_1, p_2, ..., p_m]$ be a sequence of pairs of slide and speech transcription segment, namely p_i, which complies the order constraint defined as follows:

- the speech segment of p_{i+1} is the segment following the speech segment of p_i, in a temporal point of view;
- the slide of p_{i+1} can either be the same slide as the one of p_i or the slide following the slide of p_i.

Let $\beta = \{\alpha_1, \alpha_2, ..., \alpha_l\}$ be the set of all the possible sequences α_i that respect the aforementioned constraints.

6.3.3. Transcription to slide alignment decision

Our objective is to find the best sequence among Π, that respects slide and speech segment order. The global decision process consists in choosing the sequence $\bar{H}$ which maximizes the global score obtained by the fusion (intersection) of *similarity between slide and speech transcription* and *slide and speech segment order constraint.* The sequence $\bar{H}$ is computed by using the following equation:

$$\bar{H} = arg \max_{\pi_x} (\Pi \cap \beta) \qquad (2)$$

6.3.4. Results

This section reports the experimental results made on manual, adapted and non-adapted LM transcriptions with the evaluation metrics.

6.3.4..1 Evaluation metrics Accuracy (see Equation 3) is the standard metric used to evaluate the performance of the transcription to slide alignment task (Yamamoto et al., 2003; Lu et al., 2014). The metrics looked at the alignment task as a classification problem.

$$Accuracy = \frac{Number\ of\ TS\ assigned\ to\ a\ correct\ slide}{Total\ number\ of\ TS} \qquad (3)$$

where TS is the number of transcription segments.

The metric presents the limitation of equally penalising near and far alignment decision. Indeed, an hypothesis alignment to a slide produced by the system is only considered as correct if it coincides exactly to the actual slide. Since one of the educational application goals is to facilitate navigation across transcription and slides, the accuracy

[5] https://gitlab.lif.univ-mrs.fr/benoit.favre/macaon

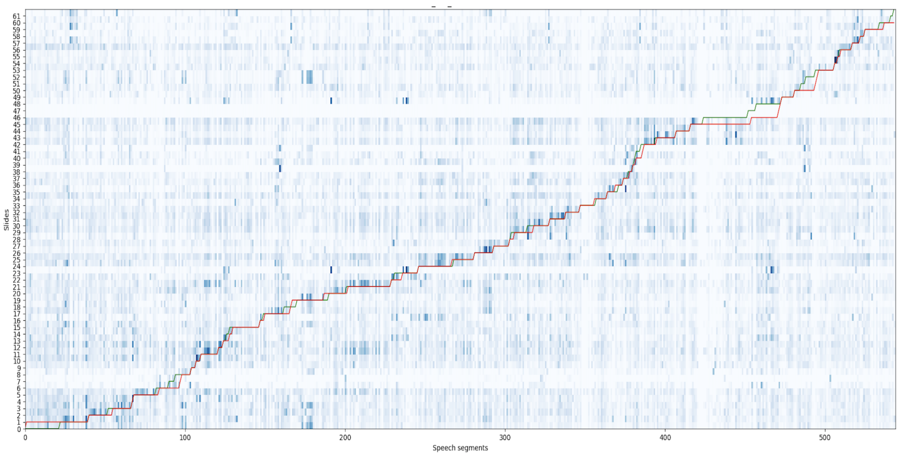

Figure 5: Alignment example for the lecture *”Introduction to algorithms”*.

Table 7: Alignment results (MSE and Accuracy) between slides and automatic transcriptions using the baseline system (Jung et al., 2018) with manual transcriptions, and automatic transcriptions without (ASR Generic) and with LM adaptation (ASR Adapt).

	ASR Generic	**ASR Adapt**	**Manual Transcription**
Accuracy	18.96%	21.49%	24.98%
MSE	681.31	638.024	657.980

Table 8: Alignment results (MSE and Accuracy) between slides and automatic transcriptions using our proposed system with manual transcriptions, and automatic transcriptions without (ASR Generic) and with LM adaptation (ASR Adapt). Text slides and speech segments are considered as two distinct collection to build the TF-IDF representation.

	ASR Generic	**ASR Adapt**	**Manual Transcription**
Accuracy	44.32%	58.46%	63.28%
MSE	2.481	1.424	1.313

Table 9: Alignment results (MSE and Accuracy) between slides and automatic transcriptions using our proposed system with manual transcriptions, and automatic transcriptions without (ASR Generic) and with LM adaptation (ASR Adapt). Text slides and speech segments are considered as one collection to build the TF-IDF representation.

	ASR Generic	**ASR Adapt**	**Manual Transcription**
Accuracy	41.19%	56.11%	62.64%
MSE	3.268	1.708	0.911

measure is not suitable because a small or a big misalignment are considered false at the same cost. As a consequence, we prefer to use the Mean Square Error (MSE) that takes into consideration the distance between the reference and the hypothesis (see Equation 4).

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(\bar{S}_i - S_i)^2 \qquad (4)$$

where m is the total number of transcription segments, $\bar{S}_i$ is the hypothesis slide assigned to a transcript segment i, and S_i is the reference slide assigned to the transcription segment i.

6.3.4..2 Baseline system The baseline system consists in a simple classification system such as the system presented in (Jung et al., 2018). The classification seeks to select for each transcription segment the slide having the highest cosine similarity. Accuracy and MSE results of the baseline system are presented in Table 7.

6.3.4..3 Experimental results We present in this section some experimental results of the proposed method. Table 8 shows the alignment performance using manual transcription, automatic adapted transcription (ASR Adapt), or automatic non-adapted transcription (ASR Generic). Line 1 and 2 illustrate respectively the performance of our system in terms of accuracy and MSE. Results show that the proposed approach obtains significant improvements compared to the baseline system. The proposed method improves the accuracy from 24.98% to 63.28% and the MSE from 658.980 to 1.313 using the manual transcription.

Experimental results reported in Table 8 shows also the usefulness of language model adaptation for the transcription to slides alignment task.

We have seen in our experimental framework (Table 4) that the automatic adaptation of LM for speech recognition allows us to reduce the global relative WER by 15.6% (WER from 19.46% to 16.4%). These values, although interesting, do not highlight the impact related to the target tasks for which the automatic transcriptions are generated (transcription to slide alignment task in our case). In terms of accuracy, the global relative reduction achieves 24.18% (accuracy from 58.46% to 44.32%) and in terms of MSE the global relative reduction achieves 42.60% (MSE from 2.481 to 1.424).

Table 9 shows the alignment performance using manual transcription by considering text slides and speech segments as one collection to build the TF-IDF representation. These results shows the usefulness of our proposed method by considering slides and speech segments as distinct collections for the automatic transcription. We observe a lost from 1.424 to 1.707 on MSE and from 58.46% to 56.11% on accuracy using the adapted LM.

Figure 5 shows an alignment example using our proposed method. The green line presents the reference and the red line presents the output of our system. Each blue square represent the TF-IDF similarity between a slide i and a speech segment j.

7. Conclusion

The paper presents the PASTEL corpus, a new French multimodal corpus containing a wide range of educational oral lectures manually transcribed and thematically segmented. In addition to the speech, the corpus contains the written presentation supports (slides) and the video. The dataset will be distributed under an open-source licence to the scientific community. We have presented some possible use cases (linguistic adaptation of speech transcription systems, thematic segmentation of oral lectures, and temporal alignment between speech and slides), including baseline results, to stimulate research community on these problematics. While these case studies presented in this article focus on the oral and written modalities, we think that the corpus finds its interest for works in image and video processing, but also in multimodality.

As future work, we plan to propose other original approaches, such as using the alignment to improve the performance of the speech recognition system by rescoring the n-best hypothesis of transcription using information from aligned slides.

8. Acknowledgments

Support for this research was provided by the french national research agency, under the PASTEL project (ANR-16-CE33-0007).

9. Bibliographical References

Akita, Y., Tong, Y., and Kawahara, T. (2015). Language model adaptation for academic lectures using character recognition result of presentation slides. In *Proc. ICASSP*.

Aubert, O. and Jaeger, J. (2014). Annotating video with open educational resources in a flipped classroom scenario. *arXiv preprint arXiv:1412.1780*.

Aubert, O., Prié, Y., and Canellas, C. (2014). Leveraging video annotations in video-based e-learning. *arXiv preprint arXiv:1404.4607*.

Auer, E., Russel, A., Sloetjes, H., Wittenburg, P., Schreer, O., Masnieri, S., Schneider, D., and Tschöpel, S. (2010). Elan as flexible annotation framework for sound and image processing detectors. In *Seventh conference on International Language Resources and Evaluation [LREC 2010]*, pages 890–893. European Language Resources Association (ELRA).

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication*, 33(1-2):5–22.

Bazillon, T., Esteve, Y., and Luzzati, D. (2008). Manual vs assisted transcription of prepared and spontaneous speech. In *LREC*.

Bettenfeld, V., Mdhaffar, S., Choquet, C., and Piau-Toffolon, C. (2018). Instrumentation of classrooms using synchronous speech transcription. In *European Conference on Technology Enhanced Learning*, pages 648–651. Springer.

Bettenfeld, V., Mdhaffar, S., Piau-Toffolon, C., and Choquet, C. (2019). Instrumentation of learning situation using automated speech transcription: A prototyping approach. In *11th International Conf on Computer Supported Education (CSEDU 2019)*.

Bouchekif, A., Damnati, G., Esteve, Y., Charlet, D., and Camelin, N. (2015). Diachronic semantic cohesion for topic segmentation of tv broadcast news. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Bouchekif, A., Charlet, D., Damnati, G., Camelin, N., and Estève, Y. (2017). Evaluating automatic topic segmentation as a segment retrieval task.

Broux, P.-A., Desnous, F., Larcher, A., Petitrenaud, S., Carrive, J., and Meignier, S. (2018). S4d: Speaker diarization toolkit in python. In *Interspeech 2018*.

Canellas, C., Aubert, O., and Prié, Y. (2015). Prise de note collaborative en vue d'une tâche: une étude exploratoire avec coconotes live.

Furui, S., Maekawa, K., and Isahara, H. (2000). A japanese national project on spontaneous speech corpus and processing technology. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*.

Furui, S., Iwano, K., Hori, C., Shinozaki, T., Saito, Y., and Tamura, S. (2001). Ubiquitous speech processing. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 13–16. IEEE.

Gravier, G., Bonastre, J.-F., Geoffrois, E., Galliano, S., McTait, K., and Choukri, K. (2004). The ESTER evaluation campaign for the rich transcription of french broadcast news. In *LREC*.

Hearst, M. A. (1997). Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational linguistics*, 23(1):33–64.

Jung, H., Shin, H. V., and Kim, J. (2018). Dynamicslide: Exploring the design space of reference-based interaction techniques for slide-based lecture videos. In *Pro-*

ceedings of the 2018 Workshop on Multimedia for Accessible Human Computer Interface, pages 33–41. ACM.

Kawahara, T., Nemoto, Y., and Akita, Y. (2008). Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. ICASSP*.

Lamel, L., Adda, G., Bilinski, E., and Gauvain, J.-L. (2005). Transcribing lectures and seminars. In *Ninth European Conference on Speech Communication and Technology*.

Lu, H., Shen, S.-s., Shiang, S.-R., Lee, H.-y., and Lee, L.-s. (2014). Alignment of spoken utterances with slide content for easier learning with recorded lectures using structured support vector machine (svm). In *Fifteenth Annual Conference of the International Speech Communication Association*.

Mdhaffar, S., Estève, Y., Hernandez, N., Laurent, A., Dufour, R., and Quiniou, S. (2019). Qualitative evaluation of ASR adaptation in a lecture context: Application to the pastel corpus. *Proc. Interspeech 2019*, pages 569–573.

Miranda, J., Neto, J. P., and Black, A. W. (2013). Improving asr by integrating lecture audio and slides. In *Proc. ICASSP*.

Mougard, H., Riou, M., De La Higuera, C., Quiniou, S., and Aubert, O. (2015). The paper or the video: Why choose? In *International World Wide Web Conference (WWW'2015)*. ACM Press.

Nasr, A., Béchet, F., Rey, J.-F., Favre, B., and Le Roux, J. (2011). Macaon: An nlp tool suite for processing word lattices. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations*, pages 86–91. Association for Computational Linguistics.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *ASRU*.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., Wang, Y., and Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Interspeech*, pages 2751–2755.

Riedhammer, K., Gropp, M., Bocklet, T., Hönig, F., Nöth, E., and Steidl, S. (2013). Lmelectures: A multimedia corpus of academic spoken english. In *First Workshop on Speech, Language and Audio in Multimedia*.

Rousseau, A., Boulianne, G., Deléglise, P., Estève, Y., Gupta, V., and Meignier, S. (2014). LIUM and CRIM ASR system combination for the REPERE evaluation campaign. In *International Conference on Text, Speech, and Dialogue*. Springer.

Trancoso, I., Nunes, R., and Neves, L. (2006). Classroom lecture recognition. In *International Workshop on Computational Processing of the Portuguese Language*, pages 190–199. Springer.

Trancoso, I., Martins, R., Moniz, H., Mata, A. I., and Viana, M. C. (2008). The lectra corpus–classroom lecture transcriptions in european portuguese. *Economic Theory*, 1(17):15–1.

Yamamoto, N., Ogata, J., and Ariki, Y. (2003). Topic segmentation and retrieval system for lecture videos based on spontaneous speech recognition. In *Eighth European Conference on Speech Communication and Technology*.

Yamazaki, H., Iwano, K., Shinoda, K., Furui, S., and Yokota, H. (2007). Dynamic language model adaptation using presentation slides for lecture speech recognition. In *Eighth Annual Conference of the International Speech Communication Association*.