UNIVERSITAT POLITÈCNICA DE VALÈNCIA
DEPARTAMENTO DE SISTEMAS INFORMÁTICOS Y COMPUTACIÓN

UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Evaluation of innovative computer-assisted transcription and translation strategies for video lecture repositories

Doctoral Thesis

presented by Juan Daniel Valor Miró
supervised by Dr. Jorge Civera Saiz
Dr. Alfons Juan Ciscar

21st August, 2017

Evaluation of innovative computer-assisted transcription and translation strategies for video lecture repositories

Juan Daniel Valor Miró

Thesis performed under the supervision of doctors
Jorge Civera Saiz and Alfons Juan Ciscar
and presented at the Universitat Politècnica de València
in partial fulfilment of the requirements for the degree
Doctor en Informática

Valencia,
21st August, 2017

The work leading to this invention has received funding from the European Union Seventh Framework Programme (FP7/2007- 2013) under grant agreement no 287755 (transLectures). Also, it has received funding from the EU's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme under grant agreement no 621030 (EMMA). In addition, it has been supported by the Spanish MINECO/FEDER research projects TIN2009-14511 (iTrans2), TIN2012-31723 (Active2Trans) and TIN2015-68326-R (MORE); and the Generalitat Valenciana under the Gerónimo Forteza program (FPA/2012/014).

Thanks to everyone who have helped me during the course of this thesis.

ABSTRACT

Nowadays, the technology enhanced learning area has experienced a strong growth with many new learning approaches like blended learning, flip teaching, massive open online courses, and open educational resources to complement face-to-face lectures. Specifically, video lectures are fast becoming an everyday educational resource in higher education for all of these new learning approaches, and they are being incorporated into existing university curricula around the world.

Transcriptions and translations can improve the utility of these audiovisual assets, but rarely are present due to a lack of cost-effective solutions to do so. Lecture searchability, accessibility to people with impairments, translatability for foreign students, plagiarism detection, content recommendation, note-taking, and discovery of content-related videos are examples of advantages of the presence of transcriptions.

For this reason, the aim of this thesis is to test in real-life case studies ways to obtain multilingual captions for video lectures in a cost-effective way by using state-of-the-art automatic speech recognition and machine translation techniques. Also, we explore interaction protocols to review these automatic transcriptions and translations, because unfortunately automatic subtitles are not error-free. In addition, we take a step further into multilingualism by extending our findings and evaluation to several languages. Finally, the outcomes of this thesis have been applied to thousands of video lectures in European universities and institutions.

RESUMEN

Hoy en día, el área del aprendizaje mejorado por la tecnología ha experimentado un fuerte crecimiento con muchos nuevos enfoques de aprendizaje como el aprendizaje combinado, la clase inversa, los cursos masivos abiertos en línea, y nuevos recursos educativos abiertos para complementar las clases presenciales. En concreto, los videos docentes se están convirtiendo rápidamente en un recurso educativo cotidiano en la educación superior para todos estos nuevos enfoques de aprendizaje, y se están incorporando a los planes de estudios universitarios existentes en todo el mundo.

Las transcripciones y las traducciones pueden mejorar la utilidad de estos recursos audiovisuales, pero rara vez están presentes debido a la falta de soluciones rentables para hacerlo. La búsqueda de y en los videos, la accesibilidad a personas con impedimentos, la traducción para estudiantes extranjeros, la detección de plagios, la recomendación de contenido, la toma de notas y el descubrimiento de videos relacionados son ejemplos de las ventajas de la presencia de transcripciones.

Por esta razón, el objetivo de esta tesis es probar en casos de estudio de la vida real las formas de obtener subtítulos multilingües para videos docentes de una manera rentable, mediante el uso de técnicas avanzadas de reconocimiento automático de voz y de traducción automática. Además, exploramos diferentes modelos de interacción para revisar estas transcripciones y traducciones automáticas, pues desafortunadamente los subtítulos automáticos no están libres de errores. Además, damos un paso más en el multilingüismo extendiendo nuestros hallazgos y evaluaciones a muchos idiomas. Por último, destacar que los resultados de esta tesis se han aplicado a miles de vídeos docentes en universidades e instituciones europeas.

RESUM

Hui en dia, l'àrea d'aprenentatge millorat per la tecnologia ha experimentat un fort creixement, amb molts nous enfocaments d'aprenentatge com l'aprenentatge combinat, la classe inversa, els cursos massius oberts en línia i nous recursos educatius oberts per tal de complementar les classes presencials. En concret, els vídeos docents s'estan convertint ràpidament en un recurs educatiu quotidià en l'educació superior per a tots aquests nous enfocaments d'aprenentatge i estan incorporant-se als plans d'estudi universitari existents arreu del món.

Les transcripcions i les traduccions poden millorar la utilitat d'aquests recursos audiovisuals, però rara vegada estan presents a causa de la falta de solucions rendibles per fer-ho. La cerca de i als vídeos, l'accessibilitat a persones amb impediments, la traducció per estudiants estrangers, la detecció de plagi, la recomanació de contingut, la presa de notes i el descobriment de vídeos relacionats són un exemple dels avantatges de la presència de transcripcions.

Per aquesta raó, l'objectiu d'aquesta tesi és provar en casos d'estudi de la vida real les formes d'obtenir subtítols multilingües per a vídeos docents d'una manera rendible, mitjançant l'ús de tècniques avançades de reconeixement automàtic de veu i de traducció automàtica. A més a més, s'exploren diferents models d'interacció per a revisar aquestes transcripcions i traduccions automàtiques, puix malauradament els subtítols automàtics no estan lliures d'errades. A més, es fa un pas més en el multilingüisme estenent els nostres descobriments i avaluacions a molts idiomes. Per últim, destacar que els resultats d'aquesta tesi s'han aplicat a milers de vídeos docents en universitats i institucions europees.

CONTENTS

CHAPTER 1

INTRODUCTION

Contents

1.1 Motivation and document structure

One of the most active research fields nowadays is Artificial Intelligence (AI), which pursues algorithms with the ability to mimic the human intelligence and give the machine the capabilities to process the natural language and to simulate the human perception. In fact, Pattern Recognition (PR) and Machine Learning (ML) are sub-fields that study how to infer specific knowledge from the data to identify patterns. Some applications of this field are Natural Language Processing (NLP), Automatic Speech Recognition (ASR) and Machine Translation (MT). These applications can be applied to the automatic generation of transcriptions and translations of educational videos, which is the focus of this thesis.

On the other hand, in the Technology Enhanced Learning (TEL) area, video lectures are widely used, not only as a complement of face-to-face [26] lectures, but also in new educational approaches like blended learning and Massive Open Online Courses (MOOCs). The adoption of video lectures in higher education is a widespread phenomenon [5] that is changing the landscape of formative options not only at universities, making lecturers think out of the box [26, 29, 37], but also at other institutions and private companies that understand video lectures as a possibility to train their personnel at low cost.

As mentioned above, video lectures are an important ingredient in MOOCs. MOOCs are rapidly growing since 2011, with more than 35 million students and 4000 courses offered at the beginning of 2016, roughly doubling the figures of the previous year [1]. Although US-based providers like edX and Coursera are now targeting international students, most courses are just delivered in English (75%), Spanish (9%), French (6%) or Chinese (4%) [2]. Clearly, for MOOCs to reach a worldwide audience, they need to be provided in multilingual form. And this also holds true for Open Educational Resources (OER) in general. Apart from its application to MOOCs and OER, multilingualism is of great interest in all contexts where educational videos are used. This includes online education in general [5, 14], flipped teaching [6, 31], and in-class recording services [16].

Although MOOCs and OER comprise objects of different kinds, in this thesis we focus our attention on producing multilingual video lectures; that is, on adding subtitles in their source spoken language and then translate them into different *target* languages. In fact, the utility of these audiovisual assets could be further improved by adding subtitles that can be exploited to incorporate added-value functionalities, even if the source subtitles are of moderate quality. This functionalities include but are not limited to searchability, accessibility, translatability, note-taking [11], improving accessibility for hearing-impaired and foreign students [7, ch.7], [24], video-clip search based on keywords [8, 25], and discovery of content-related videos [11, 18]. ASR and MT techniques can automatically generate subtitles in multiple languages for these valuable educational contents.

A direct approach to obtain source video subtitles is to generate automatic transcriptions by using ASR technology. Indeed, the application of ASR technology to lecture recordings is by no means new. A detailed account of significant efforts in this domain up to 2010 can be found in [7, ch.7]. More recent research efforts on ASR applied to educational videos

can be found in the European projects transLectures [4] and European Multiple MOOC Aggregator (EMMA) [3]. The ASR technology has reached a level of maturity that allows us to generate low-cost, automatic source subtitles of (nearly) publishable quality in most cases. It is worth noting, however, that such quality is only achievable by developing state-of-the-art ASR systems adapted to the particular task (media repository) at hand, using *massive adaptation*[a] techniques. However due to the nature of these techniques, subtitles are not error-free. For this reason, lecturers need to manually review video subtitles to guarantee the absence of errors. This thesis aims to develop and test cost-effective solutions with real-users to generate transcriptions and translations.

As with source subtitles, a straightforward approximation to obtain target video subtitles is to generate automatic translations by using MT technology. This approach has been also explored with good results in the European projects transLectures and EMMA, and more recently in TraMOOC [17]. The translation quality of adapted MT systems is often worth of post-edition; that is, it is often the case that the automatic translation is not far from the correct translation, and thus it is more time-efficient to review it than producing the entire translation manually. In addition, as in ASR, system adaptation has been shown to be a key factor in maximising output quality. It goes without saying that MT is normally applied to clean, post-edited automatic transcriptions and, as indicated above, automatic translations are also post-edited to end up with target subtitles of publishable quality. Regarding this, it is worth noting that many approaches have been considered to increase user productivity when reviewing subtitles, like the *intelligent interaction*[b] [27] approach, but post-editing is still the most popular [9, 20–22, 30, 35]. Both approaches will be assessed as part of this thesis.

Given the discussion above, in Chapter 3 we consider the integration of a state-of-the-art Spanish ASR system into the Opencast Matterhorn [15] platform, a free, open-source platform to support the management of educational audio and video content. The ASR system was trained on a novel large speech corpus, known as poliMedia [32], that was manually transcribed for the European project transLectures [34]. This speech recognition system was developed within the framework of the European *transLectures* project [34], along the lines of other systems, such as KALDI [23], JANUS-II [36], UPC RAMSES [19] or SPHINX-II [13]. Initial results on the poliMedia corpus are also reported to compare the performance of different ASR systems based on speaker and topic adaptation. Notable improvements over the baseline performance were reported, as a result of these adaptations.

As explained before, automatic subtitles need to be reviewed and post-edited in order to ensure that what students see on-screen is of an acceptable quality. So in Chapter 4 we investigate different user interface design strategies for this post-editing task to discover the best way to incorporate automatic transcription technologies into large educational video repositories. We setup a three-phase study involving lecturers from the Universitat Politècnica

[a] The process whereby automatic subtitling systems can be adapted to the lecture in question using lecture-specific material such as presentation slides, related documents, or the speaker voice.

[b] The process whereby, in the subsequent post-editing stage, automatic subtitling systems direct the user to those subtitles that contain the most transcription errors.

de València (UPV) with videos available on the UPV media repository, which is currently over 20K video objects. This three-phase study involved conventional post-editing, a transcription review strategy based on Confidence Measures (CM) and a third strategy resulting from the combination of that based on CM with massive adaptation techniques for ASR.

The next step taken in Chapter 5 is multilingualism, because as mentioned before, one of our final objectives is to reach a wider audience in MOOCs and OER by lowering the language barrier with the minimum effort. This is assessed in two real environments: the UPV media repository of video lectures, and the European EMMA MOOC platform. In fact, MOOCs and OER are not usually offered in multiple languages due to the lack of cost-effective solutions, but the previous adaptation can be extended and used to provide multilingual MOOCs and OER cost-effectively. To this purpose, ASR and MT systems for a wide range of languages were generated, including Italian and Portuguese ASR systems, and MT systems for Italian-English, English-Italian, Portuguese-English, and Dutch-English; in the framework of the transLectures and EMMA projects [3].

In Chapter 6 multilingual subtitles automatically generated by the combination of ASR and MT systems required a manual review to reach publishable quality. So, we performed a comprehensive evaluation in efficiency terms of the process of delivering multilingual video subtitles for real-life MOOCs and related OER repositories, comparing them with mainstream providers of this technology as YouTube automatic captioning system and Google Translate.

We encourage a sequential reading of all the chapters, starting by Chapter 2 in which we briefly explain some preliminary and background concepts on the research fields that will be used in the rest of chapters. However, only specific chapters can be read attending to the dependency graph shown in Figure 1.1. As can be observed, this thesis has two main branches that can be read independently: the monolingual and the multilingual. Finally, Chapter 7 gives a brief summary of the work described, highlighting the scientific publications that support this thesis.

1.2 Scientific and technological goals

Summarising the foregoing, we list below the main scientific and technological goals pursued in this thesis.

- Study how massive adaptation techniques can lead to better transcription quality (Chapter 3).
- Compare evaluation protocols to minimise user review time (Chapter 4).
- Study dependencies between transcription and translation quality and time invested in review (Chapters 4 and 6).
- Develop high-quality efficient ASR and MT systems for multiple languages (Chapter 5).
- Scientific evaluation of topic and speaker adaptations for ASR and MT (Chapter 5).

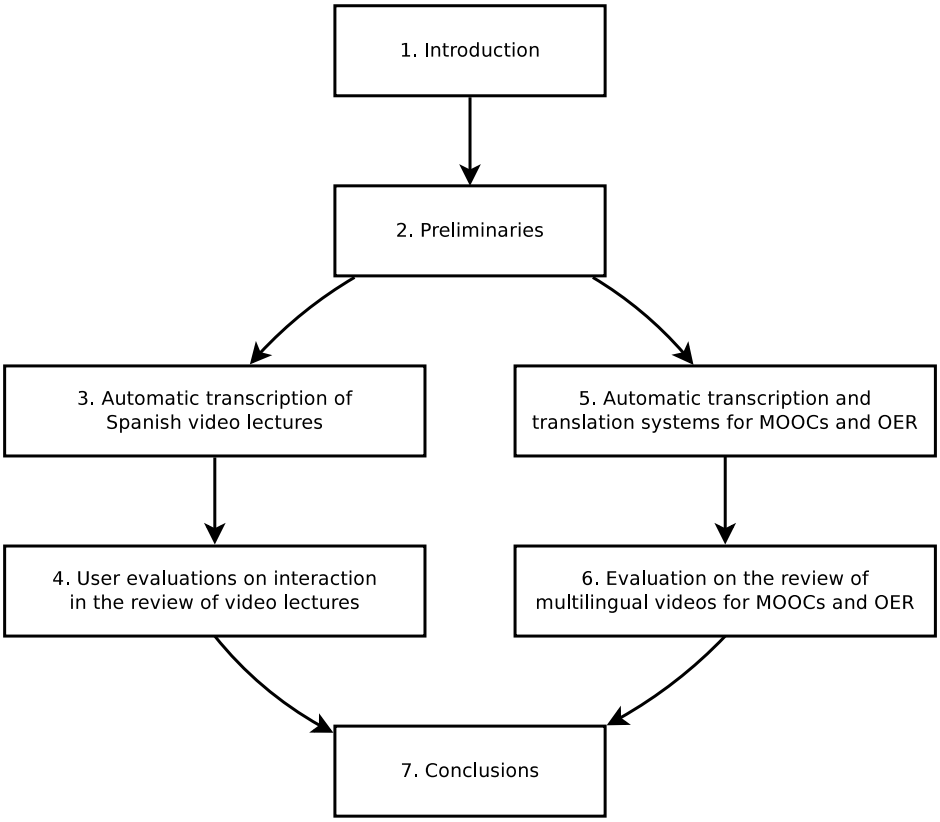

Figure 1.1: Thesis' chapter dependency graph.

- Propose and evaluate a real-life solution to enable users to edit multilingual captions (Chapter 6).

1.3 Context of this thesis

This thesis has greatly benefited from the context in which it has been developed, having access to real-life challenges provided by educational institutions. From 2011 to 2014, the UPV coordinated the European project transLectures [28] to implement automatic transcription and translation systems for video lectures based on cost-effective techniques such as, massive adaptation and intelligent interaction. transLectures tries to give an answer to the need for transcriptions and translations of video lectures [8, 10] in educational institutions and universities. To this end, two pilots were considered: the UPV media repository and the well-known videoLectures.NET repository.

Also, EMMA was a European project from February 2014 to July 2016 under the CIP-ICT PSP programme intending to form part of the EU strategy to modernise ICT-based learning and scale up higher education to meet Europe 2020 targets. The project duration was 30 months, divided into three periods of 10 months (referred as P1, P2, P3). This project pursues to showcase excellence in innovative teaching methodologies and learning approaches. This is achieved through a large-scale piloting of multilingual MOOCs on different subjects in their own MOOC platform. EMMA provides a system to deliver open online courses in multiple languages to any European university; in order to preserve Europe's rich cultural, educational and linguistic heritage, by promoting real cross-cultural and multilingual online learning.

To this purpose, during the EMMA project the so-called EMMA platform has been developed to include four special features to improve the impact in education and cover the specific needs of citizens across Europe:

- Aggregator: Any institution can add their own MOOC, or promote MOOCs held on their own platform, for free. This has offered learners a large selection of courses in a single place.
- Learning analytics: A variety of data relating to learner profile, behaviour, response and success has been collected and analysed to better understand learning processes in an online environment.
- Personal Learning Environment (PLE): A built-in PLE would allow learners to construct their own learning pathways using units from different MOOCs as building blocks.
- Multilingual: Built-in automated transcription and translation for all video lectures and text contents has allowed learners to access and understand MOOCs that are not in their mother tongue.

The aim of these two projects is to produce innovative, cost-effective technologies for the transcription and translation of the vast online collections of video lectures currently emerging in education. Video lectures are being used by universities around the world to enhance, supplement and even revolutionise traditional university curricula. The rest of the section introduces two real-life case studies in which multilingual video subtitles will be delivered as a result of this thesis: EMMA MOOCs and the UPV media repository.

1.3.1 EMMA MOOCs

During the EMMA project 12 initial partners collaborate to provide more than 20 MOOCs with their experience and expertise in the field of e-learning, learning analytics, and innovative translation technology. This platform also works as an aggregator, in order to allow any institution add their own MOOCs. A screenshot of the EMMA platform advertising some MOOCs can be observed in Figure 1.3.

In order to make accessible all the EMMA MOOCs to people with disabilities and audiences of different countries and languages, the EMMA consortium has made use of our

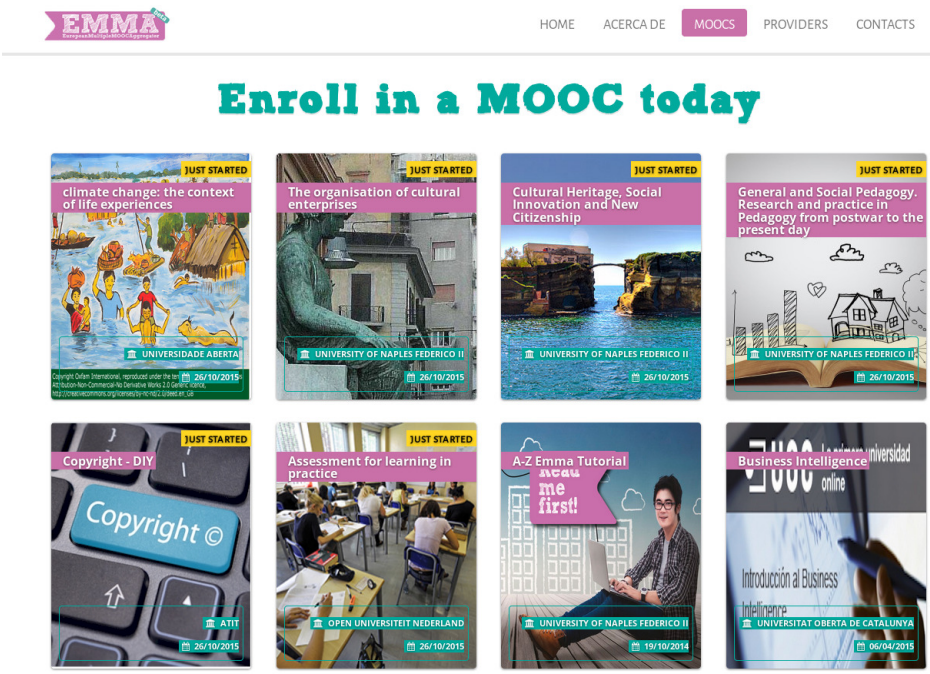

Figure 1.2: Screenshot of the EMMA platform advertising some MOOCs.

technology to generate automatic transcriptions and translations of all video (and text) contents included in the MOOCs. Providing MOOCs and OER in multilingual form means translating objects of different kinds, most notably videos and text documents. Figure 1.3 shows a unit of the trilingual MOOC Open Wine University, originally in French and translated into English and Italian. As observed, a *translation* button allows to switch between languages, and subtitles are displayed in the selected language (English).

This latter feature supported automatic video transcription in 7 languages (English, Italian, Spanish, Dutch, French, Portuguese and Estonian) and automatic video and text translation into English, and from English into Spanish and Italian. Automatic transcriptions and translations were reviewed by lecturers to reach publishable quality for the students to follow the course. Most courses were offered in the original language plus English, and English MOOCs were also provided in Spanish. It is important to note that five of these courses were delivered in three languages (English, Italian, and Spanish or French).

Table 1.1 shows the number of videos and duration (in hours) for each language which has been included in MOOCs delivered on the EMMA platform. The average duration of videos for all languages except for Dutch is less than 10 minutes. Dutch videos last more than 35 minutes on average and the format of the video presentation is different from that in the other languages. Dutch videos are interviews with usually two speakers sitting around a table, while

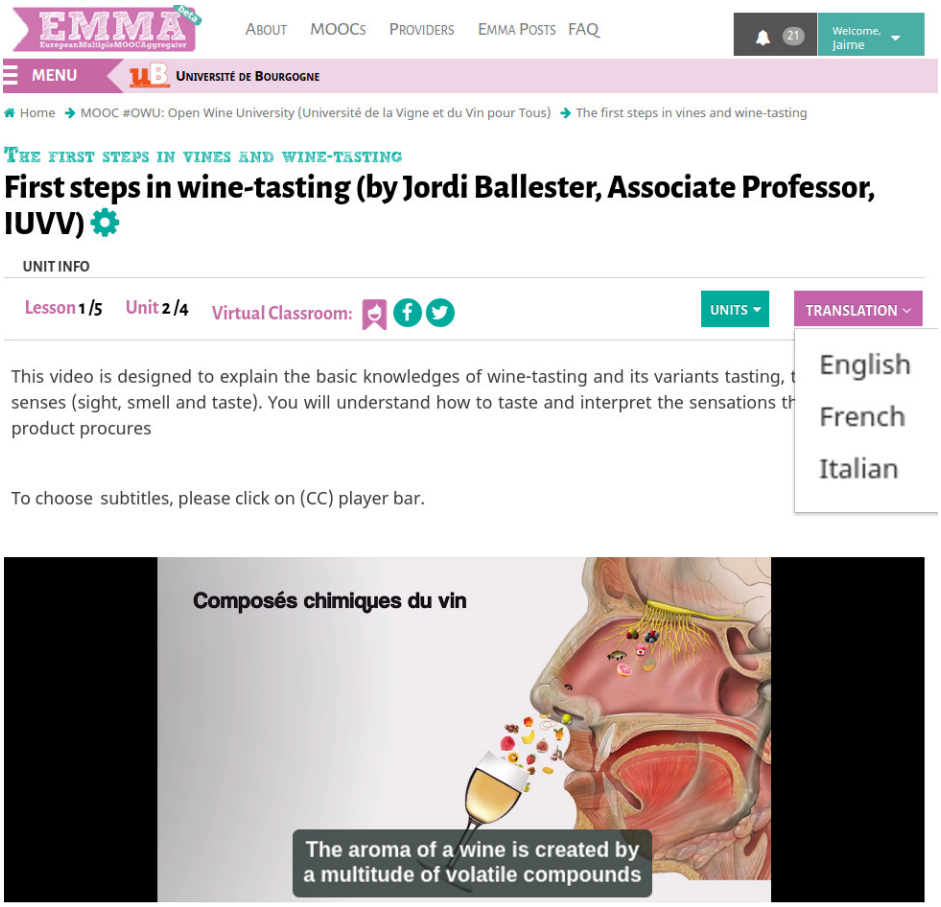

Figure 1.3: Screenshot of the trilingual MOOC # Open Wine University.

in the other videos a single speaker stands in front of the camera.

Table 1.1: Videos and duration (in hours) for each language in EMMA MOOCs.

Language	Videos	Hours
Dutch	56	34.6
Spanish	231	30.8
Italian	163	19.3
Portuguese	13	7.7
French	64	7.3
English	25	3.5
Estonian	21	1.2

1.3.2 The UPV media repository

The UPV media repository is a service available at the UPV for the creation, storage, management and dissemination of video lectures in a professional setting. This service was launched in 2007, and was designed to allow UPV lecturers to record video lectures in order to supplement traditional face-to-face classes. Nowadays, there are two kinds of media objects in the repository: poliTubes and poliMedias. poliTubes are educational videos produced by students and lecturers themselves and uploaded to the media repository in a similar fashion to YouTube. poliMedias provide a concise overview of a given topic and have a typical duration of around ten minutes [33]. These short video lectures, in fact, are the most extended video format in MOOCs, since viewers' attention rapidly drops after the first minutes being watched [12]. Given the suitability of this latter type of video, in this thesis we will focus only on them. Table 1.2 shows the statistics of the UPV media repository.

Table 1.2: Statistics of the UPV media repository.

Language	Videos	Hours	Lecturers
poliTube	19834	4165	1584
poliMedia	16875	2994	1819

Each poliMedia video focuses on a specific topic with an average duration of 10 minutes approximately. An example of a poliMedia lecture can be seen in Figure 1.4. This repository hosts more than 15.000 video lectures covering different topics, in part incentivised by the Docència en Xarxa (DeX) action plan. poliMedia is primarily designed to allow UPV lecturers to record pre-prepared mini lectures for use by students in supplement to the traditional live lecture. For the most part they consist of concise overviews of a given topic and have a typical duration of around ten minutes.

As of July 2016, 1819 lecturers have recorded more than 17000 video lectures (2934 hours), with levels of participation gaining momentum year on year since 2007. In Table 1.3 shows basic statistics on poliMedias at the UPV media repository for the three most common languages. As shown, 90% of the poliMedias are in Spanish, followed by English and Catalan. It is also worth noting the large number of lecturers involved in the recording of poliMedias.

Table 1.3: Number of poliMedia hours of video for each language.

Language	Videos	Hours	Lecturers
Spanish	15013	2709	1572
English	1221	173	203
Catalan	434	52	80

The production process for poliMedia repositories was carefully designed to achieve both a high rate of production and an output quality comparable to that of a television production, but

Figure 1.4: A video lecture host at the poliMedia platform.

at a lower cost. A poliMedia studio consists of a 4x4 metre room with a white backdrop, video camera, capture station, pocket microphone, lighting and AV equipment including a video mixer and audio noise gate. The hardware cost of this studio stands at around 15,000 euros. We should note that the reduced size of the set means we can obtain a sharper image more easily than if in a standard lecture theatre. Figure 1.5 shows a picture taken at the poliMedia recording studio during a recording session.

The recording process for poliMedia is quite simple: university lecturers are invited to come to the studio with their presentation and slides. They stand in front of the white backdrop and deliver their lecture, while they and their computer screen (presentation slides) are recorded on two different video streams. The two streams are stacked side-by-side in real-time to generate a raw preview of the poliMedia content, which can be reviewed by the lecturer at any time. These streams are then post-processed; they are cropped, joined (with some overlap) and h264 encoded to generate an mp4 file, which can be distributed online via a streaming server. All of this is fully-automated and the lecturer can review the post-processed video in a matter of minutes. The resulting video lectures have a resolution of 1280x720. Finally, the video lecture is upload to the poliMedia website, and distributed through various social and educational channels. In addition to the presentation slides, lecturers are requested to provide any metadata and additional textual resources related to the subject of the video lecture.

Lecturers at the UPV volunteer to review automatic transcriptions and translations, in most cases, of their own videos. However, it is also possible that other lecturers different from the author of the video, or even students review the subtitles of a video. In both cases, editions carried out by other users must be approved by the author.

Figure 1.5: A typical poliMedia recording session at the UPV.

Bibliography

[1] Class central. `www.class-central.com/report/moocs-2015-stats`, 2016.

[2] Class central: Languages. `www.class-central.com/languages`, 2016.

[3] European Multiple MOOC Aggregator (EMMA) project. `platform.europeanmoocs.eu`, 2016.

[4] Transcription and Translation of Video Lectures (transLectures) project. `translectures.eu`, 2016.

[5] I. E. Allen and J. Seaman. *Class differences: Online education in the United States, 2010*. Babson Survey Research Group, 2010.

[6] J. Bishop and M. A. Verleger. The flipped classroom: A survey of the research. In *Proc. of ASEE Annual Conference*, 2013.

[7] P. O. de Pablos, J. Zhao, and R. Tennyson, editors. *Technology Enhanced Learning for People with Disabilities: Approaches and Applications: Approaches and Applications*. Information Science Reference, 2011.

[8] C. Dufour, E. G. Toms, J. Lewis, and R. Baecker. User strategies for handling information tasks in webcasts. In *Proc. of CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1343–1346, 2005.

[9] M. Federico, A. Cattelan, and M. Trombetti. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proc. of AMTA*, 2012.

[10] A. Fujii, K. Itou, and T. Ishikawa. Lodem: A system for on-demand video lectures. *Speech Communication*, 48(5):516 – 531, 2006.

[11] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. of Interspeech*, 2007.

[12] P. J. Guo, J. Kim, and R. Rubin. How video production affects student engagement: An empirical study of mooc videos. In *Proc. of Learning at Scale*, pages 41–50, 2014.

[13] X. Huang, F. Alleva, H. wuen Hon, M. yuh Hwang, and R. Rosenfeld. The SPHINX-II Speech Recognition System: An Overview. *Computer, Speech and Language*, 7:137–148, 1992.

[14] R. H. Kay. Exploring the use of video podcasts in education: A comprehensive review of the literature. *Computers in Human Behavior*, 28(3):820 – 831, 2012.

[15] M. Ketterl, O. A. Schulte, and A. Hochman. Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia. In *ISM 2009, 11th IEEE International Symposium on Multimedia, San Diego, California, USA, December 14-16, 2009*, pages 687–692, 2009.

[16] M. Ketterl, O. A. Schulte, and A. Hochman. Opencast matterhorn: A community-driven open source software project for producing, managing, and distributing academic video. *Interactive Technology and Smart Education*, 7(3):168–180, 2010.

[17] V. Kordoni, A. Bosch, K. Kermanidis, V. Sosoni, K. Cholakov, I. Hendrickx, M. Huck, and A. Way. Enhancing Access to Online Education: Quality Machine Translation of MOOC Content. In *Proc. of LREC 2016*, pages 16–22, 5.

[18] T. Mei, B. Yang, X.-S. Hua, L. Yang, S.-Q. Yang, and S. Li. Videoreach: An online video recommendation system. In *Proc. of SIGIR*, pages 767–768. ACM, 2007.

[19] A. Nogueiras, J. A. R. Fonollosa, A. Bonafonte, and J. B. Mariño. RAMSES: El sistema de reconocimiento del habla continua y gran vocabulario desarrollado por la UPC. In *VIII Jornadas de I+D en Telecomunicaciones*, pages 399–408, 1998.

[20] S. O'Brien. Towards predicting post-editing productivity. *Machine Translation*, 25(3):197–215, Sept. 2011.

[21] S. O'Brien and M. Simard, editors. *Special Issue: Post-Editing*, volume 28. Machine Translation, 2014.

[22] M. Plitt and F. Masselot. A productivity test of statistical machine translation postediting in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16, 2010.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011. IEEE Catalog No.: CFP11SRW-USB.

[24] R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J. P. Robinson, and B. S. Duerstock. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, 6(4):299–311, 2013.

[25] S. Repp, A. Groß, and C. Meinel. Browsing within lecture videos based on the chain index of speech transcription. *TLT*, 1(3):145–156, 2008.

[26] T. Ross and P. Bell. "No significant difference" only on the surface. *International Journal of Instructional Technology and Distance Learning*, 4(7):3–13, 2007.

[27] I. Sanchez-Cortina, N. Serrano, A. Sanchis, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proc. of ACM IUI*, pages 325–326, 2012.

[28] J. A. Silvestre, M. Del Agua Teba, G. Gascó, A. Giménez, A. Martínez-Villaronga, I. Sanchez-Cortina, N. Serrano Martinez-Santos, **J. D. Valor Miró**, J. Andrés, J. Civera, et al. transLectures. In *IberSPEECH 2012-VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSPEECH 2012)*, pages 345–351. Universidad Autónoma de Madrid, 2012.

[29] S. K. A. Soong, L. K. Chan, C. Cheers, and C. Hu. Impact of video recorded lectures among students. *Who's learning*, pages 789–793, 2006.

[30] L. Specia. Exploiting objective annotations for measuring translation post-editing effort. In *Proc. of EAMT*, pages 73–80, 2011.

[31] B. Tucker. The flipped classroom. *Education next*, 12(1), 2012.

[32] C. Turró, M. Ferrando, J. Busquets, and A. Cañero. Polimedia: a system for successful video e-learning. In *Eunis 2009 international conference*, 2009.

[33] C. Turró, M. Ferrando, J. Busquets, and A. Cañero. Polimedia: a system for successful video e-learning. In *Eunis 2009 international conference*, 2009.

[34] UPVLC, XEROX, JSI-K4A, RWTH, EML, and DDS. transLectures: Transcription and Translation of Video Lectures. In *Proc. of EAMT*, page 204, 2012.

[35] **J**. **D**. **Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74:65–75, 2015.

[36] P. Zhan, K. Ries, M. Gavalda, D. Gates, A. Lavie, and A. Waibel. JANUS-II: towards spontaneous Spanish speech recognition. 4:2285–2288, 1996.

[37] D. Zhang, L. Zhou, R. O. Briggs, and J. F. N. Jr. Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information and Management*, 43(1):15 – 27, 2006.

CHAPTER 2

PRELIMINARIES

Contents

2.1 Introduction

The conventional approach to ASR consists in learning reasonable acoustic and language models to find the most probable transcription given an acoustic signal of, for instance, a video lecture. Usually, acoustic models are learnt from a few hundred hours of transcribed speech data, while language models require as much text data as possible. Also, MT systems are based on the so-called statistical approach to MT, proposed in the early nineties by ASR researchers willing to test statistical models in MT, but generalizing the systems by log-linearly combining different models to better fit the characteristics of the translation task [17].

The rest of this chapter is organised as follows. First, in Section 2.2 we will explain the estimation and usage of language models, that will be used in both ASR and MT systems. Then, the basics of ASR systems, training and recognition included, will be presented in Section 2.3. Then, we will do the same with the MT systems in section 2.4. Finally, in Section 2.5 we give details on the evaluation measures used in this thesis.

2.2 Language modelling

Language modelling is a probabilistic approximation to the modelization of grammatical, semantic and syntactic relations between words of a given vocabulary or language. A Language Model (LM) measures how likely is a certain sequence of words to be written in a particular language. The typical implementation of a LM is the n-gram model, which estimates the probability of consecutive groups of up to n words [6], but other approaches based on neural networks have been proposed. Next, we describe the two main techniques that we used to create state-of-the-art LMs: the linear mixture LM and the Recurrent Neural Network Language Model (RNNLM).

Linear mixture LM

The linear mixture LM [5, 11, 18, 19, 21] aims at alleviating the problem of Out-of-Vocabulary (OoV) words in large-scale vocabulary tasks with a great variety of topics [3]. Individual LMs are first trained for each main resource separately, and then combined using a linear mixture optimised on textual content extracted from the domain of our application, what is also generally referred to as in-domain text. Actually, the LM training has two steps, first the vocabulary selection, and second, the proper LM estimation.

Vocabulary selection is an important step, before estimating LMs. The higher the number of words in the vocabulary is, the larger the quantity of data to reliably estimate the LM is needed. However, when a small vocabulary is selected, words that are not included in it will not be recognised by the ASR system. These words are known as OoV words. Therefore, a trade-off between the size of the vocabulary and the quantity of data available to train the LM has to be reached.

Words to be included in the vocabulary are selected according to their probability in a unigram LM model. This unigram model is obtained as follows. First, a unigram LM is trained for each individual resource. Then, all individual unigram LM are combined using a linear interpolation optimised on in-domain text. Last, given the resulting LM model, all words are sorted by their probability, from highest to lowest, and the first n words are selected to be included in the vocabulary. A vocabulary of 50 thousand words was found to be optimal for most languages, as we will explain later. With this quantity, the ratio of OoVs is 1% on average, but including more words in the vocabulary do not significantly decrease this ratio.

Once the vocabulary is created, the LM can be estimated. This process is very similar to the vocabulary selection process, but only those words in the vocabulary are used to estimate the LM. In this case, the LM is not a simple unigram LM, but a 4-gram smoothed LM using modified Knesser-Ney discount [6] is trained for each individual resource. As in the unigram LM, individual LMs are combined by a linear interpolation optimized on in-domain text. Finally, for efficiency, n-grams below a certain probability threshold are pruned out and the final LM is obtained.

It must be noted that the training of the LM have been performed with the SRILM toolkit [30], which is freely available for non-commercial purposes.

Recurrent neural network LM

Deep Neural Networks (DNNs) have revolutionised most of the PR fields in recent years, and so, they have also been considered for language modeling. LMs based on Neural Networks are not new in ASR [4]. However, their potential improvement was limited by the expensive estimation process, and as the n-gram based model, it also failed to model long word dependencies. Notwithstanding, Mikolov [22] recently proposed an efficient version of RNNLMs applied to language modelling. This proposal basically consists in feeding the neural network not only with the current input, but also with the previous state of the network. In this way, it theoretically enables the RNNLM to model long-term dependencies between words, and practically, they have been shown to outperform n-gram based LM in terms of perplexity. However, RNNLMs are limited in terms of size and cannot be estimated from large collections of data. Consequently, they are combined together with n-gram LMs to improve ASR performance.

The RNNLM is trained using the RNNLM toolkit [23] on a small quantity of text related to the topic. Once this RNNLM is estimated, it is employed to search better transcriptions or translations.

2.3 Automatic speech recognition

Conventional ASR systems are based on the Bayes decision rule that combines two basic models: an Acoustic Model (AM) and a LM. Given an observed acoustic signal X and a hypothesis on the sentence uttered W, the acoustic model $P(X|W)$ measures how likely the

signal is to be an acoustic realization of the hypothesis, whereas the language model $P(W)$ computes a prior probability of the hypothesis to be uttered. Finally, we search for the best hypothesis $\hat{W}$ of the sentence uttered. Eq. 2.1 represents the recognition process.

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|X) = \underset{W}{\operatorname{argmax}} P(X|W)P(W) \tag{2.1}$$

Next, we describe in detail the generation of the acoustic model and the recognition process involved in ASR. In addition, we briefly describe the lexical model responsible for transforming the AM phonemes into the words of the LM.

2.3.1 Acoustic modelling

State-of-the-art AM is performed by consecutive steps. First, we generate a basic Hidden Markov Model with emiting probabilities modelled by Gaussian Mixture Models (HMM/GMM), in a sequence of steps of increasing complexity. Then, we adapt the model using the well-known speaker adaptation technique called Constrained Maximum Likelihood Linear Regression (CMLLR). The next step is to replace the HMM/GMM with a Deep Neural Network (HMM/DNN) using the alignment generated by the HMM/GMM model. This model can be further improved by using sequence-discriminative training. Finally, we can enrich our AM by using data from different languages creating a Multilingual Deep Neural Network (mDNN). Below, we explain in detail these models.

HMM/GMM acoustic models

The AM employed for all languages corresponds to a HMM/GMM model trained using the data available. This data corresponds to phonetically annotated speech segments. For instance, video lectures along with their subtitles. The training process of this model is composed by multiple steps, and for each step several techniques are employed.

First, speech files are preprocessed to reduce the noise and variability of the signal. More precisely, speech files are converted into Mel-Frequency Cepstral Coefficients (MFCCs) feature vectors of 49 dimensions [34]. MFCCs are a numerical representation of the speech signal in order to ease the recognition process. Then, mean and variance normalisation of the feature vectors is performed per each cluster defined in the training dataset. A cluster refers to a group of samples, which typically correspond to the samples of a certain speaker. This normalization helps the system to be better estimated from different speakers.

Then, given the feature vectors extracted from the speech signal and their transcription, a standard HMM/GMM model is trained from all samples, resulting in a language-independent model [25]. First, a 3-state HMM/GMM, in which the emission probability is modeled as a Gaussian mixture, is trained for each phoneme. Then, this model is refined using the Classification and Regression Tree (CART) algorithm [26]. Specifically, this algorithm employs expert knowledge of the phonetic features of human speech (vowels, nasal vowels, fricative consonants, etc.) to refine the model.

CMLLR speaker adaptation

Once this standard model is trained, a speaker-adapted model can be estimated. The standard model is used to align the speech signal and the transcription at the phoneme level for all samples in the training set. Given an aligned training set, a transformation matrix for the samples of each speaker is computed using CMLLR [9], followed by the projection of the samples of each speaker with its corresponding transformation matrix, generating speaker-independent samples.

Basically, this projection normalises speech samples of each speaker to be more homogeneous compared to other speakers. These normalised samples are used to train a CMLLR model following a similar procedure to that of the standard HMM/GMM training. At this point, two models have been estimated, one from the original speech samples, and another from the CMLLR normalised speech samples. These two models are the basis of the acoustic model of our ASR system.

HMM/DNN acoustic models

The incorporation of DNN [14] for acoustic modelling has led to huge improvements in ASR [8]. Using the HMM/GMM acoustic model, we compute an alignment between the acoustic data and the training transcriptions, to obtain a mapping from a frame of the acoustic signal to the HMM/GMM state associated with this frame [8]. Then, we train a Hidden Markov Model emiting with Deep Neural Networks (HMM/DNN) so that it predicts the model state given the acoustic data. In this way, we can compute a probability distribution over the model states for each frame. Note that in some systems the number of states is quite large. For this reason, for practical purposes the states are clustered and only the probability distribution over the clustered states is computed.

DNN sequence-discriminative training

DNN parameter estimation is usually carried out by maximising the Cross-Entropy (CE) criterion [7]. This is a cost-effective and reliable criterion defined at frame level, where each HMM/DNN state is interpreted as a class label. In this criterion, labels "compete" against each other without taking into account the transitions, the LM, and how words are transcribed according to the lexicon model. To overcome this issue, recent works have proposed to replace the CE criterion by the Maximum Mutual Information (MMI) criterion applied at utterance level, that is usually referred as sequence training for HMM/DNN [32].

The main drawback of this criterion is the computational cost required to estimate it, making this criterion unfeasible in practice. To overcome this issue, this calculation is usually approached using lattices. At the beginning of each epoch in the training of HMM/DNN, lattices are generated for all training utterances using the model from the previous epoch. These lattices are then used to approximate the denominator using the forward-backward algorithm.

Regarding the algorithm used to maximise the criterion, there are two algorithms which are commonly used: Stochastic Gradient Descent (SGD) and Resilient Backpropagation (RPROP) [27, 28]. The SGD for sequence training converges to an optimal solution faster than RPROP, but it is more unstable than RPROP. To overcome this problem, two smoothing techniques are usually employed: the frame-rejection heuristic and CE smoothing. In the first one, those frames in which the posterior probability for the ground truth label is less than a threshold are discarded. In the CE smoothing, the MMI criterion for sequence training is interpolated with the conventional CE criterion. It is worth noting, that these smoothing techniques can also be used with the RPROP algorithm.

Multilingual deep neural networks

Some languages posses a large quantity of freely annotated resources, ranging from annotated speech to electronic text. This makes easy to reliably estimate a robust ASR system and improve system performance [12]. However, there are languages in which annotated speech data is not so abundant. One way to tackle this scarcity problem in the case of speech data is to take advantage from the fact that human languages share phonemes, that is, there are speech sounds that are identical or very similar across languages.

The idea behind mDNN is to pool speech data from similar languages to robustly train acoustic models for a specific language [31]. The inner structure of this mDNN is maintained for all languages, while the output layer is estimated separately for each different language. This way it is supposed that the inner layers of the mDNN capture the speech sound, while the output layer models language-specific phonemes.

2.3.2 Lexical modelling

A lexical model provides the information of how each word in a language should be pronounced. Obviously, the lexical model to be employed in an ASR system depends on the phonetics of the language under study. This results in two types of lexical models. On one hand, non-ambiguous languages at the phonetic level possess a unique phonetic transcription for each word. In this case, phonetic rules are based on a set of simple rules such as in the case of Spanish or Italian. On the other hand, in ambiguous languages such as English or Dutch, the same word might be pronounced in different ways and simple pronunciation rules are not available. In the latter case, in order to generate the phonetic transcription of each word, a statistical grapheme-to-phoneme model [1] is trained. This model infers the phonetic transcription of new words from a limited set of phonetically annotated words. We will only refer to the lexical model if the phonetics of the language is ambiguous. Otherwise, this model is not necessary as phonetic transcription is performed using well-defined rules.

2.3.3 The recognition process

The recognition process is identical to the training process in its preprocessing step, converting video lectures into MFCCs feature vectors [34]. Then, speech data go through a 3-pass recognition process to obtain the final transcription. At each recognition pass, a LM, that is described in Section 2.2, is involved.

First, given the preprocessed speech samples, the first recognition pass is performed. Specifically, the HMM/GMM acoustic model explained in Section 2.3.1 is used to recognise the video. This results in an initial automatic transcription generated by the standard system.

Next, the second pass improves the initial transcription by performing speaker adaptation based on CMLLR [9], if available. More precisely, a CMLLR adaptation is performed using the transcription obtained in the first pass in order to obtain a set of adapted samples. Then, these samples are recognised using a CMLLR model. In this case, the result is a transcription generated by a speaker-dependent CMLLR system, which can be a HMM/GMM or HMM/DNN model.

Last, if we used a HMM/DNN model, a third pass is carried out to further better the transcriptions. In this pass, the HMM/DNN of the CMLLR model is adapted employing the transcription obtained in the second pass [35]. This adaptation is specially appealing if the quantity of in-domain data available is quite low. Once, this HMM/DNN is adapted, the adapted samples of the second pass are recognised again. Finally, the transcription obtained corresponds to the final transcription proposed by the ASR system.

If we use the RNNLM model explained in Section 2.2, the recognition is performed as explained above, but the system instead of generating only the best transcription in the last recognition-step, it generates the n-best transcriptions. These transcriptions are then re-scored by the RNNLM and that with the highest score is provided as the final transcription.

Efficient systems

Transcription systems can be tuned, achieving in most cases very competitive results. The challenge is to tune these systems for speed to be able to provide the transcription of a video in a limited amount of time without sacrificing too much transcription quality. Therefore, we can improve the efficiency of the transcription process, generating for some of our systems, an alternative efficient version.

The most straighforward way to improve the time efficiency is to make the system consider a smaller quantity of hypothesis when transcribing, in other words, prune the search space to explore it faster. However, this could turn into a significant degradation of the transcription quality if it is not carefully performed.

In order to guarantee the quality of our system, the efficient version could be generated only for those systems that already achieve high quality transcriptions, as those are the ones in which a minor loss of quality would be acceptable. Specifically, the LM can be pruned and the recognition parameters can be tuned to prune the search space aggressively. Next, the

RNNLM along with a non-pruned n-gram LM can be employed to re-score and improve the results of the pruned recognition.

Also, we can optimise the recognition process. On the one hand, the time required to load the acoustic and language models to transcribe a video can be drastically reduced by using binarized models. On the other hand, we can make a better use of the Graphical Processing Units (GPUs) involved in the recognition process.

2.4 Machine translation

Current MT systems are based on a statistical approach to the problem. There are also two basic models in this case: a language model $P(W)$, as in ASR, and a translation model $P(E|W)$. Given an input sentence E in a source language and a hypothesis W on its translation into an output sentence in a target language, the translation model measures how likely the input sentence is to be the actual sentence that leads to the given output sentence. In order to translate an input sentence, both models are combined according to the Bayes rule, to search for the best hypothesis $\hat{W}$, as shown in Eq. 2.2.

$$\hat{W} = \underset{W}{\operatorname{argmax}} P(W|E) = \underset{W}{\operatorname{argmax}} P(E|W)P(W) \tag{2.2}$$

The translation model is usually trained from parallel text data (e.g. millions of translation examples). By far, the most popular toolkit for statistical MT is Moses [16]. Next, we describe the translation modelling and the translation process in its essence. For the language model, please refer to Section 2.2.

2.4.1 Translation modelling

As mentioned above, to create the translation model representative data from the application domain is needed in order to train an effective MT system. However, this kind of in-domain translated materials are usually scarce. Fortunately, out-domain parallel texts are being continuously generated all over the world and freely available on the internet. For instance, international institutions, such as the European Commission, or educational or entertainment enterprises such as TED or OpenSubtitles provide large amounts of multilingual parallel texts. This makes easier to estimate general-purpose or out-domain MT systems. State-of-the-art MT systems employ a phrase-based approach to translate texts. Basically, the MT system is trained by extracting parallel phrases (or short sequences of consecutive words) from parallel texts and assigning a probability to them.

Intelligent selection

To solve the aforementioned domain-adaptation problem, intelligent selection techniques have been proposed to extract from out-domain parallel corpora those bilingual sentences that

would be useful to train the in-domain MT system and provide better translation quality [33]. This is specially appealing in the MOOC and OER domain, where courses to be translated correspond to specific domain content that cannot be easily translated by a general-purpose MT systems.

Intelligent selection techniques are based on similarity measures computed between the in-domain and out-domain texts. Using these measures, relevant texts from the out-domain data are extracted. Finally. the MT system will be trained on the in-domain data plus the selected set of the out-domain corpora. In the case of the MT systems we have tested four selection techniques based on three popular similarity measures: Moore [24], Axelrod [2], Mansour [20] and Infrequent *n*-gram Selection (INS) [10]. The selected out-domain parallel corpora is devoted to train a translation model using the Moses toolkit.

However, in many cases we are completely lacking of in-domain texts, due to contents highly specific and thus no parallel text for them are available. For that reason, we propose a two-step intelligent selection technique to deal with the lack of in-domain parallel text. First, a monolingual selection, Moore or INS, is employed using the text that we will translate to select a representative parallel text from the out-domain. Next, a bilingual selection, Moore, Axelrod or Mansour, is performed to select representative parallel text from out-domain again using the in-domain parallel text obtained in the previous step. Also, we can make this process automatically, to adapt our systems to new topics without human-intervention.

Efficient models

The phrase-based approach to the translation problem leads to two main challenges in terms of efficiency. First, the resulting phrase table is enormous, and so, it hardly fits in the memory of a commodity machine. Second, automatic translation requires an extensive use of this table, as a high number of possible phrases are looked up in the table each time a sentence is translated.

In order to tackle these two challenges, we have employed an advanced feature of the Moses toolkit [16], compact phrase models. Compact phrase models [15] are an intelligently compressed version of the phrase-based table, that it is optimised to be employed in the translation process. Its application is transparent, as it obtains the same automatic translations as the standard phrase table. However, in terms of speed and memory consumption compared to conventional phrase models, translations with these compact phrase models are generated 5 times faster and the memory usage is 15 times lower.

Similarly to phrase-based tables, the LM of a MT system also influences on its efficiency. Specifically, baseline systems generated by Moses, employ LMs based on the SRILM toolkit. This LMs can be substituted by LMs based on the KenLM toolkit [13]. KenLM supersedes SRILM in both speed and size of the resulting models. But no significant differences are obtained in the results, showing the adequacy of these new LMs.

2.4.2 The decoding process

Given a text to be translated and a MT system, the automatic translation process is performed as follows. First, the input text is preprocessed in the same way as the training dataset. Next, a phrase-based decoder translates the text, sentence by sentence. Basically, the decoder, first analyzes all the possible phrases in the source sentence; next, these phrases are look up in the so-called phrase-table of the MT system to return bilingual phrases; and finally, the target sentence resulting from the concatenation of the most probable sequence of bilingual phrases matching the source sentence is returned as automatic translation. In our case, this translation process is performed also by the Moses toolkit.

2.5 Evaluation measures

To carry out fast performance evaluations of our systems, we need to use automatic evaluation measures. This kind of measures are convenient because we need to adjust a lot of parameters and compare different systems, that will be used in real-life evaluations with lecturers. Lecturers cannot review all possible systems due to the great amount of time that this would require, slowing down the development process.

For this reason, based on automatic measures we can perform fast and reliable improvements to our systems, in order to obtain better models and results. As we use well-known measures that provide a reliable measurement capability, we expect to find letter a direct relationship between these measure and the lecturers review in our final systems. We use automatic measures widely used by the scientific community, that we describe below.

Word error rate

Word Error Rate (WER) is used to measure errors in automatic transcriptions. In order to compute this measure we need a reference transcription (with the correct content) and the automatic transcription from which we need to calculate the WER. WER is computed as the number of insertions (n_i), deletions (n_d) and substitutions (n_s) between these two transcriptions divided by the number of words in the reference (n_r) as we can observe in Eq. 2.3.

$$WER = \frac{n_s + n_i + n_d}{n_r} \cdot 100 \tag{2.3}$$

WER can be thought of as a percentage approximation of the number of words that need to be corrected in order to achieve the reference transcription. For example, if a lecturer needs to apply 30 elementary editing operations to an automatic transcription so as to obtain a reviewed version of 200 words in length, then the WER will be 15%.

Bilingual evaluation understudy

One of the most used measures to measure the automatic translation quality is the Bilingual Evaluation Understudy (BLEU), which computes different *n*-gram order precisions between the hypothesis and one or more references, and combine them. BLEU is computed as indicated in Eq. 2.5.

$$\text{BLEU} = \text{BP} \cdot \left(\sum_{n=1}^{N} w_n \log p_n \right)$$

BP stands for Brevity Penalty, a factor used to penalise short translations, N is the maximum n-gram order, p_n the n-gram precision of order n, and w_n the weight assigned to that n-gram precision. All in all, BLEU can be intuitively understood as the degree of overlap between the automatic translation and the correct one. The BLEU is a quality measure so, the higher value, the better translation quality. Typically, a translation with BLEU above 30 is considered to be of acceptable quality.

Translation error rate

The Translation Error Rate (TER) [29], is a measure similar to WER, but it also counts for phrasal shifts (n_p). It is expressed as a percentage, of the number of edit operations required to convert the automatic translation into the correct reviewed translation, divided by the total number of words in the reviewed translation. Thus, if a user needs to correct 20 words in a translation containing 100 words, the TER is 20. TER computation is expressed in Eq. 2.5.

$$\text{TER} = \frac{n_s + n_i + n_d + n_p}{n_r} \cdot 100$$

Real time factor

Real Time Factor (RTF) is used to measure the time spent by the lecturers in the review process. This measure takes into account the time to review the transcription or translation (P) and the duration of the video lecture (T), and it is defined as the ratio between these two values. Equation 2.4 shows the RTF computation.

$$RTF = \frac{P}{T} \tag{2.4}$$

So if, for instance, a video lasts 20 minutes and the review of its automatic transcription takes 1 hour, the RTF for this video would be 3.

Bibliography

[1] Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 50(5):434 – 451, 2008.

[2] A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proc. of*

EMNLP, Edinburgh (UK), 2011.

[3] J. R. Bellegarda. Statistical language model adaptation: review and perspectives. *Speech Communication*, 42(1):93 – 108, 2004. Adaptation Methods for Speech Recognition.

[4] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, Mar. 2003.

[5] S. Broman and M. Kurimo. Methods for combining language models in speech recognition. In *Proc. of Interspeech*, pages 1317–1320, 2005.

[6] S. F. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pages 310–318, 1996.

[7] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(1):30–42, January 2012.

[8] G. E. Dahl, S. Member, D. Yu, S. Member, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, 2012.

[9] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[10] G. Gascó, M.-A. Rocha, G. Sanchis-Trilles, J. Andrés-Ferrer, and F. Casacuberta. Does more data always yield better translations? In *Proc. of EACL*, pages 152–161, 2012.

[11] J. T. Goodman. Putting it all together: Language model combination. In *Proc. of ICASSP*, pages 1647–1650, 2000.

[12] A. Halevy, P. Norvig, and F. Pereira. The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2):8–12, Mar. 2009.

[13] K. Heafield. Kenlm: Faster and smaller language model queries. In *In Proc. of the Sixth Workshop on Statistical Machine Translation*, 2011.

[14] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, 29(6):82–97, November 2012.

[15] M. Junczys-Dowmunt. Phrasal rank-encoding: Exploiting phrase redundancy and translational relations for phrase table compression. *Prague Bull. Math. Linguistics*, 98:63–74, 2012.

[16] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, Prague (Czech Republic), 2007.

[17] P. Koehn, F. J. Och, and D. Marcu. Statistical phrase-based translation. In *Proc. of NAACL*, Edmonton (Canada), 2003.

[18] X. Liu, M. Gales, J. Hieronymous, and P. Woodland. Use of contexts in language model interpolation and adaptation. volume Proc. of Interspeech, 2009.

[19] X. Liu, M. Gales, J. Hieronymous, and P. Woodland. Language model combination and adaptation using weighted finite state transducers. 2010.

[20] S. Mansour, J. Wuebker, and H. Ney. Combining translation and language model scoring for domain-specific data filtering. In *International Workshop on Spoken Language Translation*, pages 222–229, San Francisco, California, USA, Dec. 2011.

[21] A. Martínez-Villaronga, M. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013*, pages 8450–8454, Vancouver (Canada), 2013.

[22] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur. Recurrent neural network based language model. In *INTERSPEECH 2010*, pages 1045–1048, 2010.

[23] T. Mikolov, S. Kombrink, A. Deoras, L. Burget, and J. Cernocky. Rnnlm-recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, pages 196–201, 2011.

[24] R. C. Moore and W. Lewis. Intelligent selection of language model training data. In *Proc. of ACL*, pages 220–224, 2010.

[25] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, Feb 1989.

[26] W. Reichl and W. Chou. Robust decision tree state tying for continuous speech recognition. *Speech and Audio Processing, IEEE Transactions on*, 8(5):555–566, Sep 2000.

[27] M. Riedmiller and H. Braun. A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In *Proc. of International Conference on Neural Networks*, pages 586–591, 1993.

[28] D. J. Sakrison. Efficient recursive estimation; application to estimating the parameters of a covariance function. *International Journal of Engineering Science*, 3(4):461–483, 1965.

[29] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas, 2006*, pages 223–231, 2006.

[30] A. Stolcke. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, pages 257–286, 2002.

[31] Z. Tüske, R. Schlüter, and H. Ney. Multilingual hierarchical mrasta features for asr. In *Proc. of Interspeech*, pages 2222–2226, 2013.

[32] S. Wiesler, P. Golik, R. Schluter, and H. Ney. Investigations on sequence training of neural networks. In *In Proc. of ICASSP*, pages 4565–4569, April 2015.

[33] J. Wuebker, H. Ney, A. Martínez-Villaronga, A. Giménez, , A. Juan, C. Servan, M. Dymetman, and S. Mirkin. Comparison of Data Selection Techniques for the Translation of Video Lectures. In *Proc. of the Eleventh Biennial Conf. of the Association for Machine Translation in the Americas (AMTA-2014)*, Vancouver (Canada), Oct. 2014.

[34] M. Xu, L.-Y. Duan, J. Cai, L.-T. Chia, C. Xu, and Q. Tian. Hmm-based audio keyword generation. In K. Aizawa, Y. Nakamura, and S. Satoh, editors, *Advances in Multimedia Information Processing - PCM 2004*, volume 3333 of *Lecture Notes in Computer Science*, pages 566–574. Springer Berlin Heidelberg, 2005.

[35] K. Yao, D. Yu, F. Seide, H. Su, L. Deng, and Y. Gong. Adaptation of context-dependent deep neural networks for automatic speech recognition. In *in Proc. SLT'12*, pages 366–369, 2012.

CHAPTER 3

AUTOMATIC TRANSCRIPTION OF SPANISH VIDEO LECTURES

Contents

3.1 Introduction

As explained in Chapter 1, transcription and translation of video lectures is needed to make them accessible to speakers of different languages and to people with disabilities. Automatic transcription in these domains is however a challenging task due to many factors such as unfavourable recording quality, high rate OoV words or multiplicity of speakers and accents. This chapter presents an automatic speech recognition system to provide cost-efficient solutions to produce accurate transcriptions in Spanish.

First results are reported on the initial version of the poliMedia corpus using a linear combination of language models, as explained in Section 2.2. The baseline ASR system is based on the RWTH ASR system [8, 10] and the SRILM toolkit [14], both state-of-the-art software in speech and language modeling, respectively. Also, we present two main improvements to the baseline system. First, by interpolating the baseline language model with a LM trained on the well-known *Google n-gram* dataset [9]. Second, by using the well-known CMLLR speaker adaptation technique. Furthermore, details about the integration of this speech recognition system into the open-source video lecture platform Matterhorn are also provided. The integration into Matterhorn allows us to reach a large educational audience by allowing matterhorn-based platforms to be easily integrate ASR technology.

The rest of this chapter is organised as follows. First, the novel freely available poliMedia corpus is presented in Section 3.2. Secondly, the Opencast Matterhorn platform is introduced in Section 3.4. In Section 3.3, we perform a deep experimentation in topic and speaker adaptation. Finally, conclusions are drawn and future lines of research are depicted in Section 3.5.

3.2 The poliMedia corpus

As explained in Section 1.3.2, the UPV media repository [15] is a service available at the UPV for the creation and publication of video lectures at the UPV. To provide an in-domain dataset to train, adapt and evaluate an ASR Spanish system, 704 videolectures in Spanish corresponding to 115 hours were manually transcribed using the tool Transcriber [3] (see Table 3.1). These transcribed videolectures were selected so that authors had granted open access to their content. This new corpus is called poliMedia corpus.

Most of the videos in poliMedia were annotated with topic and keywords. More precisely, 94% of the videos were assigned a topic and 83% were described with keywords. However, these topics and keywords were not derived from a thesaurus, such as EuroVoc. Speakers were also identified for each video.

3.3 Experiments on speaker and topic adaptation

Our baseline ASR system is the RWTH ASR system [8, 10] along with the SRILM toolkit [14]. The RWTH ASR system includes state-of-the-art speech recognition technology for acoustic

Table 3.1: Basic statistics on the poliMedia corpus

Videos	704
Speakers	111
Hours	115
Sentences	40K
Running words	1.1M
Vocabulary (words)	31K
Singletons (words)	13K

model training and decoding. It also includes speaker adaptation, speaker adaptive training, unsupervised training, a finite state automata library, and an efficient tree search decoder. SRILM toolkit is a widespread language modeling toolkit which have been applied to many different natural language processing applications. In this case, we train a baseline system of HMM/GMM acoustic models, as described in Section 2.3.1.

Also, we propose to improve our baseline system by incorporating external resources to enrich the baseline LM. To this purpose, we consider the linear combination of an in-domain language model, such as that trained on the poliMedia corpus, with an external large out-domain language model computed on the Google N-Gram corpus [9]. A single parameter λ governs the linear combination between the poliMedia language model and the Google N-Gram model, being optimised in terms of perplexity on a development set.

In order to study how the linear combination of language models affects the performance, in terms of WER, of an ASR system in the poliMedia corpus, a speaker-independent partition in training, development and test sets was defined. The statistics of this partition can be found in Table 3.2. Topics included in the development and test sets range from technical studies such as architecture, computer science or botany, to art studies such as law or marketing.

Table 3.2: Basic statistics on the poliMedia partition.

	Training	Development	Test
Videos	559	26	23
Speakers	71	5	5
Hours	99	3.8	3.4
Sentences	37K	1.3K	1.1K
Vocabulary	28K	4.7K	4.3K
Running words	931K	35K	31K
OOV (words)	-	4.6%	5.6%
Perplexity	-	222	235

The baseline system, including acoustic, lexicon and language models, was trained only on the poliMedia corpus. System parameters were optimised in terms of WER on the development set. A significant improvement of more than 5 points of WER was observed when moving from monophoneme to triphoneme acoustic models. Triphoneme models were inferred using the conventional CART model using 800 leaves. In addition, the rest of parameters to train the acoustic model were 2^9 components per Gaussian mixture, 4 iterations per mixture and 5 states per phoneme without repetitions. The LM was an interpolated trigram model with Kneser-Ney discount. Higher order n-gram models were also assessed, but no better performance was observed.

Provided the baseline system, a set of improvements based on the language model were proposed and evaluated. The baseline language model solely trained on poliMedia corpus was interpolated with the Google N-Gram corpus [9]. To this purpose, we unify all Google N-Gram datasets, which are initially splitted by years, in a single, large file. Then, we train a trigram LM using Google N-Gram that was interpolated with the poliMedia LM. These two LMs were interpolated to minimise perplexity on the development set. This interpolation was performed using a particular vocabulary in the case of Google N-Gram, ranging from that vocabulary matching that of poliMedia (poliMedia vocab), over the 20.000 most frequent words in the Google N-Gram corpus (20K vocab), to the 50.000 most frequent words (50K vocab). In this latter experiment, approximate values of interpolation weights are 0.65 for the poliMedia LM and 0.35 for the Google N-Gram LM.

Next, we improved our ASR system by replacing the RWTH toolkit by the transLectures-UPV Toolkit (TLK) [4], which consists of a set of tools that allows acoustic model training and speech decoding. This toolkit was developed under the transLectures European project [12], and uses the ASR state-of-the-art techniques. Besides, as in the RWTH system, the SRILM toolkit [13] is used to estimated *n*-gram language models. More precisely, a Spanish ASR system based on a tied triphoneme HMM/GMM trained on the poliMedia corpus was deployed. In addition, the well-known CMLLR [5] technique for speaker adaptation, explained in Section 2.3.1, was applied to our system. This leads to a two-step recognition process, as CMLLR adaptation is performed using the transcription obtained in the first pass in order to obtain a set of adapted samples, that will be better recognised in the second one. The LM was a linear mixture trained on the poliMedia transcriptions along with other external resources, not only Google N-Gram. Table 3.3 summarises their main statistics.

In fact, an interpolated 4-gram language model was trained, smoothed with modified Kneser-Ney [7]. As in our interpolation with Google n-gram alone, we limit the final LM vocabulary to 50K words. The idea behind these experimental setups was to evaluate the effects, in terms of WER, of an increasing vocabulary coverage using external resources in the presence of a comparatively small in-domain corpus such as poliMedia. Experimental results are shown in Table 3.4.

As reported in Table 3.4, there is a significant improvement of 5.7 points of WER over the baseline when considering a language model trained with the 50K most frequent words in the Google N-Gram corpus. As expected, the decrease in WER is directly correlated with the

Table 3.3: Basic statistics of corpora used to generate the LM

Corpus	Sentences	Running words	Vocabulary
EPPS	132K	0.9M	27K
news-commentary	183K	4.6M	174K
TED	316K	2.3M	133K
UnitedNations	448K	10.8M	234K
Europarl-v7	2 123K	54.9M	439K
El Periódico	2 695K	45.4M	916K
news (07-11)	8 627K	217.2M	2 852K
UnDoc	9 968K	318.0M	1 854K

Table 3.4: Evolution of WER above the baseline for the RWTH ASR system, as a result of interpolating the poliMedia language model with an increasingly larger vocabulary language model trained on the Google N-Gram corpus.

System	WER	OoV
RWTH baseline	39.4	5.6%
RWTH + poliMedia vocab	34.6	5.6%
RWTH + 20K vocab	33.9	4.4%
RWTH + 50K vocab	33.7	3.5%
TLK System	30.3	1.6%
TLK System + CMLLR	24.6	1.6%

number of OoVs in the test set, since the Google N-Gram corpus provides better vocabulary coverage.

Replacing the RWTH ASR system by TLK allows us to have more control and to fine-tune the system. As part of this fine-tuning that involves the integration of additional LMs to reduce OoV words, we observe a decrease of 3.4 WER. Furthermore, the application of speaker adaptation based on the CMLLR technique implemented in TLK decreases the WER by 5.7 points.

3.4 Integration of ASR into the Matterhorn platform

Matterhorn [6] is a free, open-source platform to support the management of educational audio and video content. Institutions will use Matterhorn to produce lecture recordings, manage existing video, serve designated distribution channels, and provide user interfaces to engage students with educational videos.

Matterhorn is an open source; this means that the product is fully based on open source products. The members of the Opencast Community have selected Java as programming language to create the necessary applications and a Service-Oriented Architecture (SOA) infrastructure. The overall application design is highly modularised and relies on the Open Services Gateway initiative (OSGi) technology. The OSGi service platform provides a standardised, component-oriented computing environment for cooperating network services.

Matterhorn is as flexible and open as possible and further extensions should not increase the overall complexity of building, maintaining and deploying the final product. To minimise the coupling of the components and third party products in the Matterhorn system, the OSGi technology provides a service-oriented architecture that enables the system to dynamically discover services for collaboration. Matterhorn uses the Apache Felix [2] implementation of the OSGi R4 Service Platform [1] to create the modular and extensible application.

One main goal in transLectures was to develop tools and models for integration of ASR technology into the Matterhorn platform that can obtain accurate transcriptions by intelligent interaction with users. For that purpose, an HTML5 media player prototype was built in order to provide a user interface to enable interactive edition and display of video transcriptions (see Figure 3.1). This prototype offers a main page where available poliMedia videolectures are listed according to some criteria. Automatic video transcriptions are obtained from the ASR system when playing a particular video.

Since automatic transcriptions are not error free, an interactive transcription editor allows intelligent user interaction to improve transcription quality. However, as users may have different preferences while watching a video, the player offers two interaction models depending on the user role: simple user and collaborative user (prosumers).

Simple users are allowed to interact in a very simplistic manner, just showing their liking about the transcriptions. However, collaborative users may provide richer feedback to correct transcriptions. As shown in Figure 3.1, collaborative users have an *edit transcription* button available on the player control bar that enables the transcription editor panel. The editor panel is situated next to the video. It basically contains the transcription text, which is shown synchronously with the video playback. Clicking on a transcription word or sentence enables the interactive content modification. User corrections are sent to the speech recognition module through a web service, so corrections are processed and new transcription hypothesis are offered back to the user. Some other user-friendly features such as keyboard shortcuts and useful editing buttons are also available. Simple users have no edit transcription button available as they are not expected to be working on transcription editing. This HTML5 prototype communicates with the ASR system through a web service implemented for that purpose. Figure 3.2 illustrates the system architecture and the communication process.

The next step was to integrate the developed interactive ASR system into the Matterhorn infrastructure. There are many different approaches to perform this integration. Our proposal lets an external system manage all the transcriptions, so there will not be necessary to add nor store them in any way into the current Matterhorn system. In addition, two primary tasks are involved in the integration process into Matterhorn. Both of them require an interface to

Figure 3.1: HTML5 player and interactive transcription editor for collaborative users.

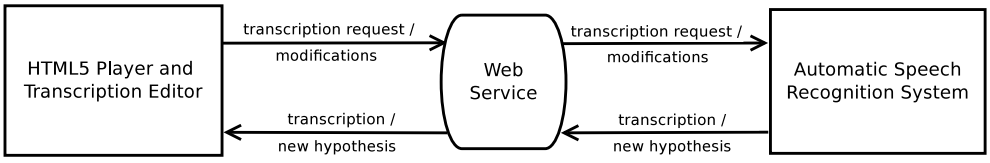

Figure 3.2: HTML5 player and ASR system communication.

enable communication between Matterhorn and the ASR system. For that purpose, a RESTful Web Service was implemented to allow media uploading, retrieve the processing status of a particular recording, request a video transcription, send transcription modifications and other functionalities.

The first task was to define a new Matterhorn workflow operation to transfer the audio data of the new media to the ASR system through the REST service mentioned before, so as to obtain automatic transcriptions for every recording uploaded to the Matterhorn platform. This task involved the implementation of a new Matterhorn service. The second part was to replace or adapt the Matterhorn Engage Player to enable transcription edition, along the lines

of the HTML5 player prototype indicated previously. The player must obtain and transmit every transcription-related information through the REST Web Service in a similar way as the HTML5 prototype did (see Figure 3.2). Here the main problem was the addition of new features to the Flash-based Matterhorn player, since it is not straightforward to implement the transcription functionalities provided by the HTML5-based player. Finally, our solution was to use an alternative open-source Matterhorn engage player based on HTML5 called Paella Player [16].

3.5 Conclusions

In this chapter we have presented a Spanish ASR system trained on a novel large speech corpus, known as poliMedia, that was manually transcribed for the European project transLectures. Then, we described the integration of this state-of-the-art ASR system into the Opencast Matterhorn platform.

Initial results on the poliMedia corpus are also provided to compare the performance of different systems. First of all, an ASR system with the RWTH toolkit was built using state-of-the-art HMM/GMM models and the in-domain poliMedia corpus for both: AM and LM training data. Then, we improved the LM using a linear interpolation with an external large-vocabulary dataset, the well-known Google N-Gram corpus. WER figures reported denote the notable improvement over the baseline performance as a result of incorporating the vast amount of data contained in the Google N-Gram corpus. Finally, we improved the interpolated LM in the TLK ASR systemperforming speaker adaptation using CMLLR.

In any case, ASR accuracy is not high enough to produce fully automatic high-quality transcriptions, and human intervention is still needed in order to reach a reasonable quality. However, user feedback can be exploited to minimise user effort in future interactions with the system [11]. For this reason, the integration into the Matterhorn platform to achieve an effective user interaction is highly valuable.

Bibliography

[1] Osgi alliance. osgi r4 service platform. http://www.osgi.org/Main/HomePage, May 2012.

[2] Apache. Apache felix. http://felix.apache.org/site/index.html, May 2012.

[3] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication special issue on Speech Annotation and Corpus Tools*, 33(1–2), 2000.

[4] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. The transLectures-UPV toolkit. In *Proc. of IberSpeech 2014*, Las Palmas de Gran Canaria (Spain), 2014.

[5] M. J. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[6] M. Ketterl, O. A. Schulte, and A. Hochman. Opencast matterhorn: A community-driven open source solution for creation, management and distribution of audio and video in academia. In *ISM 2009, 11th IEEE International Symposium on Multimedia, San Diego, California, USA, December 14-16, 2009*, pages 687–692, 2009.

[7] R. Kneser and H. Ney. Improved backing-off for m-gram language modeling. In *1995 International Conference on Acoustics, Speech, and Signal Processing, ICASSP '95, Detroit, Michigan, USA, May 08-12, 1995*, pages 181–184, 1995.

[8] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, and H. Ney. The rwth 2007 tc-star evaluation system for european english and spanish. In *Proc. of Interspeech*, pages 2145–2148, 2007.

[9] J. B. Michel et al. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

[10] D. Rybach, C. Gollan, G. Heigold, B. Hoffmeister, J. Lööf, R. Schlüter, and H. Ney. The rwth aachen university open source speech recognition system. In *Proc. of Interspeech*, pages 2111–2114, 2009.

[11] I. Sánchez-Cortina, N. Serrano, A. Sanchis, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proc. of IUI*, pages 325–326, 2012.

[12] J. A. Silvestre, M. del Agua, G. Garcés, G. Gascó, A. Giménez-Pastor, A. Martínez, A. P. G. de Martos, I. Sánchez, N. S. Martínez-Santos, R. Spencer, **J. D. Valor Miró**, J. Andrés-Ferrer, J. Civera, A. Sanchís, and A. Juan. translectures. In *Proceedings of IberSPEECH 2012*, 2012.

[13] A. Stolcke. SRILM – an extensible language modeling toolkit. In *Proc. of ICSLP*, pages 257–286, 2002.

[14] A. Stolcke. SRILM - An Extensible Language Modeling Toolkit. In *Proc. of ICSLP*, 2002.

[15] C. Turró, M. Ferrando, J. Busquets, and A. Cañero. Polimedia: a system for successful video e-learning. In *Eunis 2009 international conference*, 2009.

[16] Universitat Politècnica de València. Paella player. `http://paellaplayer.upv.es/`, 2009.

CHAPTER 4

USER EVALUATIONS ON INTERACTION IN THE REVIEW OF VIDEO LECTURES

Contents

4.1 Introduction

As discussed in Chapter 3, in the framework of the European project transLectures , automatic subtitles in Spanish were generated for all videos in the UPV media repository. However, as it stands, the quality of the automatic transcriptions generated requires lecturer intervention in order to guarantee the accuracy of the material made available to students [15]. So UPV lecturers, having filmed videos for the UPV media repository as part of an earlier DeX call, trialled the computer-assisted transcription system *transLectures player* with editing capabilities for keyboard and mouse [24].

Some previous computer-assisted transcription tools are limited to batch-oriented passive user interaction strategies in which the initial transcription is manually post-edited. More precisely, the transcription tool Transcriber has been presented together with some tests to measure the time needed to generate a transcription from scratch [1]. Some exhaustive analysis of a collaborative user post-editing system has been performed, concluding that reviewing automatic transcriptions allow to obtain useful transcriptions for educational purposes [14]. Also, it has been proved that the usage of interactive correction methods are useful for reducing WER significantly by applying speaker adaptation techniques [8]. However, none of these works assess the impact on user effort. In fact, the are some limited studies that shows a user effort reduction when transcriptions are improved with a semantic and syntactic transcription analysing tool highlighting misspelled words [20]. Finally, a batch-oriented passive user interaction protocol without system participation has been tested obtaining good results in terms of user effort, similar to those obtained in the present study [2]. However, these studies do not perform an exhaustive comparison of different user interaction methods and the relationship between quality and time devoted by the lecturer based on real-life end-user evaluations.

We adapted the computer-assisted transcription system described in Chapter 3 to serve the two main use cases that are shown in Fig. 4.1. In the first use case (on the left), lecturer recordings are automatically transcribed off-line using an ASR system. While in the second use case (on the right), users interact with a web player in order to amend recognition errors found in the automatic transcriptions previously generated.

In the first use case, our ASR system is used to generate automatic transcriptions of the video lectures. In the second use case, the user can watch and review the transcription of a video with the transLectures web player. Corrections made by the user are sent back to the web service to update the transcription file. The transLectures player interface consists of an evolution of the player presented in Section 3.4. It is an innovative web player with editing capabilities, complete with alternative display layout options and full keyboard support. This player was developed as part of transLectures at the UPV [25], in accordance with Nielsen's usability principles [17, 18]; and it was iteratively improved during subsequent evaluations described in detail in the next section.

In this chapter, we provide an in-depth analysis of a series of more intelligent active user interaction strategies for the generation of transcriptions that are accurate enough to be

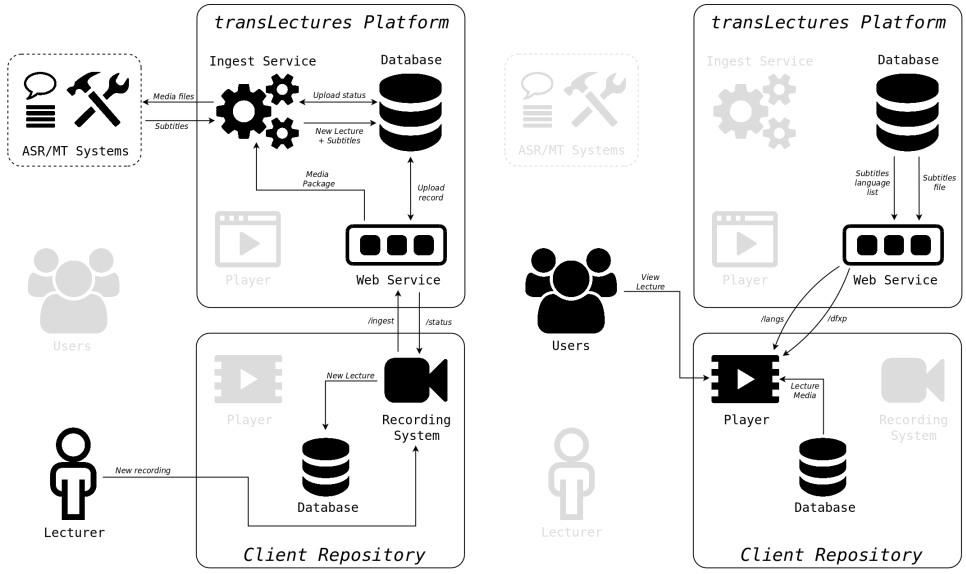

Figure 4.1: Main two use cases for video transcription (left side) and transcription revision by users (right side).

useful to students while requiring the minimum effort on the part of the lecturer [12]. To this end, a three-phase evaluation process was set up to analyse alternative user interaction strategies for reviewing the automatically-generated transcription. Our first phase consisted of a conventional manual post-editing strategy. For the second we introduced the premise of intelligent interaction, before moving onto a third phase which combines the best features from phases one and two in a two-step review process.

4.2 Methodology of the user trials

Here, we describe user evaluations carried out under UPV's DeX programme. An on-going incentive-based programme to encourage university lecturers at the UPV to develop digital learning resources based on Information and Communication Technologys (ICTs).

A total of 27 lecturers signed up for this study, reviewing a sample of 86 video lectures organised into three phases. Most participants had degrees in different branches of engineering (17), while the rest mastered business management (6), social science (2) and biology (2). Lecturers involved committed to reviewing the automatic transcriptions of five of their poliMedia videos. Lectures to be reviewed were allocated across three consecutive evaluation phases, described below.

1. Conventional post-editing: Automatic transcriptions for the first video of each lecturer are manually reviewed. Automatic transcription segments are up to 20 words long and are shown in synchrony with the video.

2. Intelligent interaction: In this phase, only a subset of probably incorrectly-recognised (low confidence) words were reviewed in the second and third videos by lecturers. These words are played within a context of one word before and one word after, being possible to expand the context to more words.

3. Two-step review: This phase organised in two consecutive rounds of evaluation for the fourth and fifth videos. The first round mimics phase two above, where the lecturer reviewed only the least confidence words. However, in this phase, least confidence words are preceded by a context of three words. Once this first round is completed, the video is then automatically re-transcribed on the basis of the lecturer's review actions preserving their corrections. In a second round, the updated transcriptions are completely reviewed as in the first phase.

Feedback from lecturers is fundamental in order to inform the design of each subsequent evaluation phase and, ultimately, of the web interface itself. The transLectures player logged precise user interaction statistics, such as the duration for which the editor window is open, the number of segments (individual subtitles) edited out of the total and the display layout selected. It also logged statistics at the segment level, including the number of mouse clicks and key presses, editing time, and the number of times a segment is played. From these statistics we computed two of the main variables of this study: RTF[a] is the time spent by the lecturer reviewing transcriptions, and WER as an indicator of the minimum number of corrections required to bring the initial automatic transcriptions into line with the reviewed transcription.

However, we also assess the impact of the three aforementioned evaluation phases in terms of WER reduction per RTF unit. That is, by how many WER points the transcription error is reduced for each RTF unit spent reviewing the automatic transcription. This ratio can be understood as a review efficiency measure, i.e. error reduction per unit of time.

In addition, feedback from lecturers was collected as subjective statistics after each phase, in the form of a brief satisfaction survey based on [11]. Lecturers were asked to rate various aspects on a Likert scale from 1-10 (see Table 4.1. They were then asked the following three open-ended questions, allowing them to freely express their subjective impressions of using the transLectures player:

- If you were to add new features to the player, what would they be?
- If you had to work with this player on a daily basis, what would you change?
- Any additional comments.

[a]In our study, the Real Time Factor (RTF) is calculated as the ratio between the time spent reviewing the transcription of a video and the duration of said video. So if, for example, a video lasts twenty minutes and its review takes, by way of example only, sixty minutes, then the RTF for this video would be 3.

The use of the satisfaction surveys over the three phases has proved to be a very valuable tool for collecting lecturers' subjective feedback and has led directly to the improvement and refinement of the transLectures player.

Table 4.1: Questions scored on a 1-10 Likert scale presented to lecturers after each phase.

Intuitiveness
1- I am satisfied with how easy it is to use this system.
2- It was easy to learn to use this system.
3- The help information of this system is clear.
4- The organisation of information on screen is clear.
Likeability
5- I feel comfortable using this system.
6- I like using the interface of this system.
7- Overall, I am satisfied with this system.
Usability
8- I can complete my work effectively using this system.
9- I can complete my work quicker than doing it from scratch.
10- This system has all the functions that I expect to have.

4.3 Experimental results

In this section we describe the experimental results attained over the three consecutive evaluation phases: conventional post-editing, intelligent interaction, and two-step review protocols.

4.3.1 First phase: Post-editing

In the first phase, 20 UPV lecturers reviewed the automatic transcription of their first video lecture in its entirety using the transLectures player, shown in Figure 4.2. A total of 2.6 hours in 20 video lectures were completely reviewed by the lecturers. Prior to this phase, lecturers were sent a link to a demo video explaining how to review their video transcriptions, in order to become familiar with the functionality of the transLectures player. The transLectures player plays the video and the transcription in synchrony, allowing the user to read the transcription while watching and listening to the video. When the lecturer finds a transcription error, it can be amended by clicking (or pressing *Enter* on the incorrect segment to pause the video. With the video paused, the lecturer can easily enter their changes in the text box that opens. Lecturers save their work periodically updating both transcription and user interaction statistics.

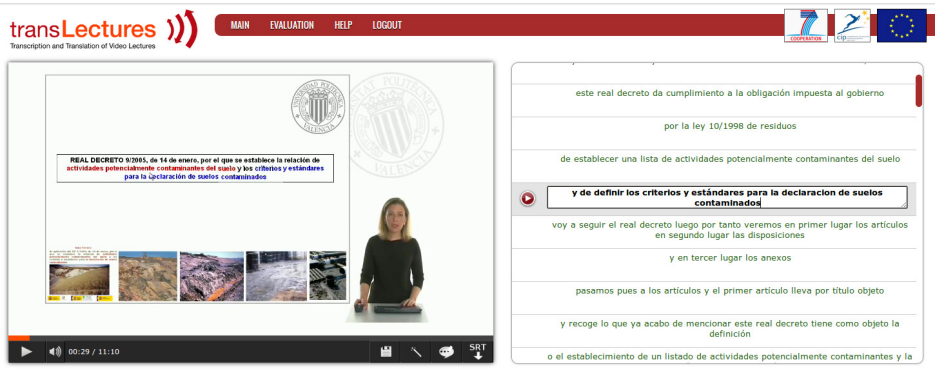

Figure 4.2: transLectures web player with the side-by-side layout while the lecturer edits one of the segments.

To assess the impact of automatic transcription on the total time required to generate usable subtitles for video lectures, we first compared times to that spent performing the same task manually from scratch. We carried out the statistical two-sample Welch's t-test for RTF with the data collected in this first phase and the data collected in the Chapter 3, in which around 100 hours of video lectures from the same repository were transcribed from scratch [25] by non-expert users (lecturers and doctoral students). We found that there was a statistically significant difference between mean RTFs (sig[b]=$5.41 \cdot 10^{-10}$), with the mean RTF for subtitles generated automatically (Mean (M)=5.4, Std (S)=2.9) being notably lower than that for those generated manually from scratch (M=10.1, S=1.8). This result suggests that the automatic transcriptions (at their reported accuracy in terms of WER) allow lecturers to generate subtitles much more efficiently than manually from scratch. We should note that the background expertise of our lecturers (engineering vs. non-engineering) was not ultimately statistically significant in terms of RTF when reviewing automatic transcriptions (sig=0.24). In addition, we also computed the WER reduction per RTF unit (M=3.2, S=1.3) to compare the effectiveness of this interaction strategy with those proposed in the second and third phases.

As shown in Table 4.2, three linear regression models were evaluated to explain RTF as a function of the independent variables of our study (WER, Intuitiveness, Likeability and Usability). Model 1 revealed that WER ($beta$[c]=0.285, sig=$4.73 \cdot 10^{-9}$) was statistically significant and accounted to a large extent for the variance observed in the data ($R^2 = 0.842$). We also considered the possibility of including the *Intercept* in this regression model, but the variance explained by the model dropped drastically.

A graphical representation of our data in terms of WER vs. RTF, and our prior knowledge of user behaviour (users essentially ignore automatic transcriptions above a certain WER

[b]It is the probability of observing an effect given that the null hypothesis is true.

[c]It is the coefficient multiplying the predictor in the linear regression model.

threshold, preferring to transcribe from scratch) suggested that a logarithmic model might better fit our data. Consequently, the logarithmic Model 2 was proposed, resulting in a more statistically significant *beta* (*beta*=2.025, *sig*=$9.82 \cdot 10^{-12}$) and an increase in the variance explained by the model ($\Delta R^2 = 0.075$).

Table 4.2: Linear regression models to explain RTF using different factors.

Predictor	*beta*	*sig*
Model 1 ($\triangle R^2 = 0.842$, $R^2 = 0.842$, *sig*=$4.73 \cdot 10^{-9}$)		
WER	0.285	$4.73 \cdot 10^{-9}$
Model 2 ($\triangle R^2 = 0.075$, $R^2 = 0.917$, *sig*=$9.82 \cdot 10^{-12}$)		
$log_e(WER)$	2.025	$9.82 \cdot 10^{-12}$
Model 3 ($\triangle R^2 = 0.001$, $R^2 = 0.918$, *sig*=$1.59 \cdot 10^{-8}$)		
$log_e(WER)$	2.263	0.007
Intuitiveness	0.144	0.832
Usability	-0.302	0.665
Likeability	0.084	0.874

As expected, both Model 1 and 2 would point that WER does in fact influence lecturer review time as expressed in RTF. Finally, we decided to incorporate the subjective variables as defined in the satisfaction survey in Table 4.1: intuitiveness (*sig*=0.832), usability (*sig*=0.665) and likeability (*sig*=0.874). However, the outcomes were ultimately not statistically significant as a means of determining RTF. This result confirms informal comments made by lecturers to the effect that transcription quality should be improved as a priority over further modifications to the user interface.

As shown in Table 4.3, lecturers felt (Overall Mean (OM) = 9.1) that the user interaction strategy in this phase was designed in accordance with intuitiveness (Grand Mean (GM) = 9.3), likeability ($GM = 8.8$) and usability ($GM = 8.9$) principles, with intuitiveness being the most highly rated characteristic.

Comments from the three open-ended questions proved to be a valuable source of feedback for refining minor usability issues and incorporating additional new features, such as changing the font size and colour, allowing the lecturer to download the transcription file being reviewed, automatically saving the transcription file and minimising the initial loading time. All in all, results were largely positive and, as desired, lecturers were able to become familiar with the transLectures player in advance of the next two phases.

Given Model 2 that is shown in Table 4.2, a more detailed user model was derived in order to predict the performance of potential user interaction strategies before being tested on real users. For the sake of interpretability, variables were expressed in absolute rather than relative terms. In other words, the independent variable WER was given in terms of word-level editing operations, while the dependent variable RTF was replaced by the time taken in seconds.

Table 4.3: Detailed results of the satisfaction survey in the first phase.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	9.4
2- It was easy to learn to use this system.	9.4
3- The help information of this system is clear.	9.2
4- The organisation of information on screen is clear.	9.0
Grand Mean	*9.3*
Likeability	
5- I feel comfortable using this system.	8.7
6- I like using the interface of this system.	8.7
7- Overall, I am satisfied with this system.	9.0
Grand Mean	*8.8*
Usability	
8- I can complete my work effectively using this system.	9.0
9- I can complete my work quicker than by doing it from scratch.	8.6
10- This system has all the functions that I expect to have.	9.0
Grand Mean	*8.9*
Overall Mean	*9.1*

As shown in Table 4.4, our statistically significant Model 1 ($R^2 = 0.801$, sig=$2.2 \cdot 10^{-16}$) correlates the time spent generating accurate subtitles with the number of correct ($beta = 1.370$, sig=$2.2 \cdot 10^{-16}$) and incorrect ($beta = 4.388$, sig=$2.2 \cdot 10^{-16}$) words in the automatic transcriptions given to our lecturers. More interesting from the point of view of the user model is the ratio between the *beta* value for independent variables (correct and incorrect words), which suggests that it takes on average three times longer to correct an incorrectly-recognised word than to confirm a correctly-recognised word.

Model 2 in Table 4.4 (R^2=0.808, sig=$2.2 \cdot 10^{-16}$) factorises the incorrect words into the three basic word edit operations: deletion ($beta$=2.059, sig=$3.2 \cdot 10^{-6}$), substitution ($beta = 4.800$, sig=$2.2 \cdot 10^{-16}$) and insertion ($beta$=5.237, sig=$2.2 \cdot 10^{-16}$), while the variable correct words ($beta$=1.370, sig=$2.2 \cdot 10^{-16}$) remains the same. The beta values can be interpreted as reflecting the relation between the time taken to perform an edit operation on an incorrect word and that taken to review a correct word, that is, essentially consisting of listening to it. As expected, simply deleting an incorrect word takes only slightly longer than reviewing a correct word. However, substitutions and insertions are more costly edit operations, requiring three to four times as long.

Defining this user model was a key step in exploring alternative, more time-effective user interaction strategies to post-editing for generating accurate subtitles for video lectures. These

Table 4.4: Linear regression on review time provided word-level edit operations.

Predictor	*beta*	*sig*
Model 1 ($\triangle R^2$=0.801, R^2=0.801, F=2030, sig=$2.2 \cdot 10^{-16}$)		
Correct Words	1.370	$2.2 \cdot 10^{-16}$
Incorrect Words	4.388	$2.2 \cdot 10^{-16}$
Model 2 ($\triangle R^2$=0.007, R^2=0.808, F=1060, sig=$2.2 \cdot 10^{-16}$)		
Correct Words	1.370	$2.2 \cdot 10^{-16}$
Deleted Words	2.059	$3.2 \cdot 10^{-6}$
Substituted Words	4.800	$2.2 \cdot 10^{-16}$
Inserted Words	5.237	$2.2 \cdot 10^{-16}$

strategies are deployed in the next two phases.

4.3.2 Second phase: Intelligent interaction

This second phase incorporates a new interaction strategy called *intelligent interaction* [23] in order to study if review times could be further improved. This strategy is based on the application of Active Learning (AL) techniques to ASR [3]. More concretely, we apply batch AL based on uncertainty sampling [10] using CM [6, 21, 26], which provide an indicator as to the probable correctness of each word appearing in the automatic transcription. In practice the lecturer may need to review (confirm) some correctly-recognised words incorrectly identified as errors (false positives), but many of the incorrectly-recognised words are spotted correctly (true positives). The idea is to focus user's review actions on incorrectly-transcribed words saving time and effort.

In this phase, lecturers are to review the subset of least confidence word according to the Computer-Aided Transcription (CAT) system in increasing order of probable correctness. This subset typically constituted between 10-20% of all words transcribed using the ASR system, though lecturers could modify this range at will to as low as 5% and as high as 40%, depending on the perceived accuracy of the transcription. Each word was played in the context of one word before and one word after, in order to facilitate its comprehension and resulting correction.

Figure 4.3 shows a screenshot of the transcription interface in this phase. Low-confidence words are shown in red and corrected low-confidence words in green. The text box including the low-confidence word can be expanded in either direction to increase the context. For this phase, the intelligent interaction mode was activated in the transLectures player by default, though lecturers could switch back to the conventional (fully manual) post-editing strategy.

Interaction statistics revealed that 12 of the 23 lecturers participating in this second phase stayed in the intelligent interaction mode for the full review of one of their poliMedia videos.

Figure 4.3: A screenshot of the transcription interface in intelligent interaction mode. Low-confidence words appear in red and reviewed low-confidence words in green. The word being edited in this example is opened for review, and the text box can be expanded to the left or right by clicking on << or >>, respectively. Clicking the green check button to the right of the text box confirms the word as correct.

In fact 2.8 hours over 18 video lectures were reviewed using that technique. In the other cases (3 hours over 22 video lectures), lecturers switched back to the conventional post-editing mode. Lecturers wanted to make sure that perfect transcriptions were obtained no matter how much time could be saved by the intelligent interaction mode. As a result, 18 videos were reviewed using intelligent interaction, while 22 videos were reviewed in the conventional post-editing mode. The RTF of the videos completely reviewed using the conventional post-editing mode (as in the first phase) was 5.2. Given the starting WER of 19.5, this time factor is comparable to results recorded in phase one.

For those lecturers that remained in the intelligent interaction mode, review time was reduced to an RTF of 2.2, though the resulting transcriptions were not error-free, unlike in phase one. That said, the residual WER of the transcriptions after being reviewed was as low as 8.0, which is not so far from that achieved by non-expert transcriptionists [7]. This indicates that confidence measures successfully identify approximately half of all incorrectly-recognised words. However, we should also assess the impact of the intelligent interaction strategy in terms of WER reduction per RTF unit. That is, by how many WER points the transcription is improved for each RTF unit spent reviewing the automatic transcription, compared to

conventional post-editing. To do so, we carried out a statistical test between intelligent interaction (M=4.6, S=3.9) and conventional post-editing (M=3.9, S=1.3). The results indicated that there was no statistically significant difference between these two strategies in this respect (*sig*=0.486). This means that intelligent interaction is in fact just as efficient in terms of WER decrease per RTF unit as conventional post-editing.

We can see in Table 4.5 that lecturers showed ($OM = 7.2$) a clear preference for obtaining perfect transcriptions, irrespective of the relative time savings afforded by the intelligent interaction strategy, and insisted on an interaction mode that gave them full control over the end quality of the transcriptions. The figures collected on intuitiveness ($GM = 8.1$), likeability ($GM = 6.8$) and usability ($GM = 6.3$), dropping from the conventional post-editing phase, reflect this assessment. However, lecturers did seem to embrace confidence measures, suggesting that low confidence words denoted in red could be incorporated into the conventional post-editing strategy.

Table 4.5: Detailed results of the satisfaction survey for intelligent interaction.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	7.8
2- It was easy to learn to use this system.	8.1
3- The help information of this system is clear.	8.1
4- The organisation of information on screen is clear.	8.4
Grand Mean	*8.1*
Likeability	
5- I feel comfortable using this system.	6.5
6- I like using the interface of this system.	6.9
7- Overall, I am satisfied with this system.	6.9
Grand Mean	*6.8*
Usability	
8- I can complete my work effectively using this system.	6.7
9- I can complete my work quicker than by doing it from scratch.	6.6
10- This system has all the functions that I expect to have.	5.6
Grand Mean	*6.3*
Overall Mean	*7.2*

User satisfaction surveys statistically reflected that post-editing ($OM = 9.1$, S=1.3) was preferred over intelligent interaction ($OM = 7.2$, S=1.7) by our lecturers (sig=$4.0 \cdot 10^{-6}$). Feedback from the three open-ended questions in the satisfaction survey clearly indicated that the intelligent interaction strategy needed rethinking in order to allow the following operations: editing of words outside of the intelligent interaction text boxes, unlimited use of the text box

expansion arrows (currently restricted to a given number of words before and after) in order to correct entire segments, and movement between text boxes in both directions (currently limited to moving forwards to the next only). Lecturer preferences notwithstanding, the intelligent interaction strategy based on confidence measures was proven to be an effective means of identifying incorrectly-recognised words. For this reason, we designed the third phase in such a way as to take greater advantage of the intelligent interaction strategy, while also granting lecturers full control over the final transcription quality.

4.3.3 Third phase: Two-step supervision

As mentioned above, the third phase was organised into two subphases or rounds and is essentially a combination of the previous two phases. In this phase, lecturers first review a subset of the least confidence words, as in the second phase. The videos are then re-transcribed (by ASR) on the basis of all previous review actions preserving those corrections made by users. These updated transcriptions are expected to be of high quality than the original transcriptions [22] reducing overall review times. In the second round of this third phase, lecturers completely review the entire re-transcription as in phase one. The fourth and fifth video of each lecturer was reviewed in this phase. Figure 4.4 shows a screenshot of the transLectures web player used in step one.

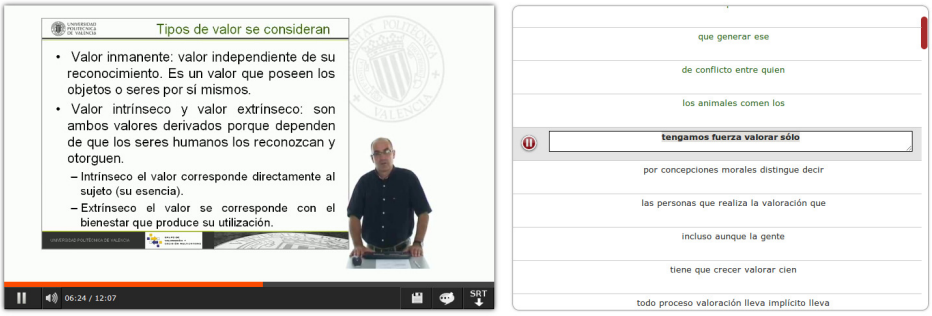

Figure 4.4: Screenshot of the transLectures web player used in step one of phase three, side-by-side layout. Each segment contains four words, of which the last word is the low-confidence word.

More concretely, the first round is devoted to review isolated segments of four words in which the last word was the low-confidence word. These segments were presented to the lecturer for review in increasing order of confidence (of the last word) until one of the following three conditions was met:

1. The total review time reached double the duration of the video itself; or

2. No corrections were entered for five consecutive segments; or

3. 20% of all words were reviewed.

The reviewed transcriptions in this phase, but also in phases one and two, were used to adapt the ASR system via a process of massive adaptation. Specifically, we adapted the acoustic models to the speaker with the Maximum Likelihood Linear Regression (MLLR) technique [5], and the language models using a linear interpolation between the language model trained on the reviewed transcriptions and the large language model previously trained [13]. Then, the automatic transcriptions were regenerated, preserving those segments already reviewed by lecturers, and using them to improve the recognition of the context words using a constrained search [9, 23]. This two-step review process was successfully completed by 15 lecturers on a total of 26 video lectures with 3.7 hours of video. More precisely, a total of 1.0 and 2.7 hours were reviewed in the first and second steps, respectively.

In the first step of this phase, average review time was as low as 1.4 RTF. As reported in Table 4.6, WER dropped significantly from the initial 28.4 to the regenerated transcriptions 18.7. That is, almost 10 WER points over 1.4 RTF, meaning that intelligent interaction plus adaptation (M=8.6, S=5.8) achieved a higher statistically significant WER reduction per RTF unit (sig=$6.9 \cdot 10^{-3}$) than intelligent interaction alone (M=4.6, S=3.9). This suggests that intelligent interaction plus adaptation is, in fact, more effective in terms of WER decrease per RTF unit than intelligent interaction alone.

In the second step, lecturers completely reviewed the regenerated transcriptions to obtain perfect final transcriptions, as in the first phase. Average RTF for this task stood at 3.9. As expected, when comparing WER reduction per RTF unit in the first phase (M=3.2, S=1.3) and the second step of this phase three (M=5.3, S=2.0), we can observe a statistically significant learning curve (sig=$8.5 \cdot 10^{-5}$) in lecturers' performance. As a result, we proved that there is a learning curve involved in getting to grips with the transLectures player.

Table 4.6: Summary of results obtained in the two-step review phase

	WER	RTF	ΔRTF
Initial transcriptions	28.4	0.0	-
First step: Intelligent interaction	25.0	1.4	1.4
Massively adapted transcriptions	18.7	1.4	-
Second step: Complete review	0.0	5.3	3.9

In order to fairly compare the first (M=3.2, S=1.3) and third (M=6.0, S=2.0) phases in terms of WER reduction per RTF unit, we subtract the effect of the learning curve for each lecturer. To this purpose, the WER reduction per RTF unit of each lecturer in the second step of this third phase was assumed to be that of the same lecturer revising their first video. This assumption leads to a corrected WER reduction per RTF unit (M=4.7, S=2.8). Even so, we found a lower yet statistically significant difference (sig=0.02) in favour of the third phase

explained by the application of massive adaptation. This result suggests that the two-step strategy is more efficient than the conventional post-editing strategy.

However, this statistically significant difference only holds when enough reviewed data is available for adaptation. That is, the reviewed data generated in the first step of this phase (aprox. 4 minutes per lecturer) is not sufficient to improve the ASR performance so that it reduces the user effort. In this latter scenario, the resulting WER after applying massive adaptation would be 24.0 instead of 18.7, resulting in a WER reduction per RTF unit (M=3.7, S=2.3) not statistically significant better (*sig*=0.31) than that obtained in the first phase. For this reason, as mentioned above, our experiments were carried out using video lectures reviewed in the previous phases, that accounted for up to approximately 25 minutes of audio data per lecturer. This amount of supervised data can be efficiently generated beforehand for each speaker using the conventional post-editing strategy in almost any real-life scenario, and then exploited in the application of a two-step supervision strategy in the subsequent videos of the same speaker.

In this phase, the best outcomes of both previous phases were successfully combined to obtain error-free end transcriptions at a lower RTF on the part of the lecturers, using a minimum amount of supervised data generated beforehand to perform massive adaptation.

Finally, note that the two-step supervision implied that lecturers have to put time aside on two separate occasions to review the same video. However, lecturers preferred to carry out the review process in a single step rather than in two steps (*sig*=0.06). This fact was reflected on the average score of the user satisfaction surveys (M=7.8, S=2.0), shown in Table 4.7. For this reason, the two-step strategy was less preferred by lecturers than the post-editing strategy.

4.4 Conclusions

In this chapter, we performed a study in the review of transcriptions and translations by lecturers, to test whether the review of automatic transcriptions was more efficient than generating them from scratch. Alternative user interaction strategies were explored to generate subtitles from automatic transcriptions as efficiently and comfortably as possible for our lecturers [16].

To this purpose, first of all, we determine that WER was the main factor involved in explaining the values of RTF. Indeed, the linear regression model derived from our data seems to generalise appropriately for transcriptions with higher WER scores than those reported here. However, it should be noted that this is a limitation of our study, since our WER figures for all video transcriptions tend to be in the range from 20 to 25. In line with [12], more sophisticated user interfaces alone, like our intelligent interaction strategy, were not proven more efficient in terms of WER decrease per RTF unit than conventional post-editing, nor were they preferred by lecturers over the simple (though more time-costly) interactive model. We find it particularly noteworthy how important it was for lecturers to be able to produce high quality (perfect) end transcriptions, prioritising this over any time-savings afforded by

Table 4.7: Detailed results from the satisfaction survey for the two-step review strategy.

Question	Mean
Intuitiveness	
1- I am satisfied with how easy it is to use this system.	7.5
2- It was easy to learn to use this system.	8.6
3- The help information of this system is clear.	8.5
4- The organisation of information on screen is clear.	8.7
Grand Mean	*8.3*
Likeability	
5- I feel comfortable using this system.	7.3
6- I like using the interface of this system.	7.4
7- Overall, I am satisfied with this system.	7.4
Grand Mean	*7.4*
Usability	
8- I can complete effectively my work using this system.	7.7
9- I can complete my work quicker than by doing it from scratch.	7.4
10- This system has all the functions that I expect to have.	7.1
Grand Mean	*7.4*
Overall Mean	*7.8*

the more intelligent strategies [4, 15, 19]: a half of our lecturers reverted to the conventional post-editing model to complete the review of their video transcriptions.

Nevertheless, the combination of intelligent interaction with massive adaptation techniques led to statistically significant savings in user effort in comparison to intelligent interaction and to the conventional post-editing strategy when sufficient adaptation data is available. This conclusion differs from that of [12] mainly because a greater amount of adaptation data has been used in our study to effectively perform the adaptation of acoustic and language models.

Our study analyses the learning curve primarily observed in the third phase as a result of lecturers having worked with the transLectures player in previous phases. WER decrease per RTF unit was statistically significantly less pronounced in the first phase than in the second step of the third phase. In this respect, Figure 4.5 shows the evolution of RTF as a function of WER across the three phases. It should be noted that the data points (video transcription reviews) of the second phase correspond to those lecturers that declined to use intelligent interaction and switch back to the conventional post-editing strategy. Data points of the third phase in Figure 4.5 are those obtained in the second step of that phase. As observed in the linear adjustment to the data points at each phase, as lecturers gain experience at reviewing transcriptions, their RTF figures improve phase-on-phase.

Furthermore, our study reveals statistically significant savings in user effort in the two-step

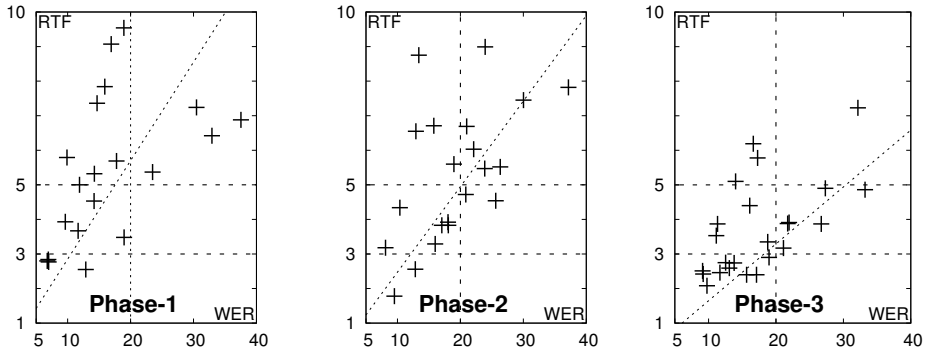

Figure 4.5: Evolution of RTF as a function of WER in the post-editing mode across the three phases. Data points of the second phase correspond to those lecturers that declined to use intelligent interaction and switch back to the conventional post-editing strategy. Data points of the third phase are those obtained in the second step of that phase.

strategy when compared to the post-editing strategy of the first phase. Intelligent interaction plus massive adaptation as a preliminary step brought significant improvements in WER to the table, that cannot solely be explained by the effect of learning curve. All in all, to our surprise, lecturers preferred the simple "one-step" post-editing strategy over the sophisticated two-step strategy.

Bibliography

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

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

[3] L. Deng and X. Li. Machine learning paradigms for speech recognition: An overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(5):1060–1089, May 2013.

[4] B. Favre, K. Cheung, S. Kazemian, A. Lee, Y. Liu, C. Munteanu, A. Nenkova, D. Ochei, G. Penn, S. Tratz, et al. Automatic human utility evaluation of ASR systems: does WER really predict performance? In *Proc. of Interspeech*, pages 3463–3467, 2013.

[5] M. J. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[6] D. Hakkani-Tur, G. Riccardi, and A. Gorin. Active learning for automatic speech recognition. In *Proc. of ICASSP*, volume 4, pages 3904–3907, 2002.

[7] T. J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proc. of Interspeech 2006*, pages 1606–1609, 2006.

[8] H. Kolkhorst, K. Kilgour, S. Stüker, and A. Waibel. Evaluation of interactive user corrections for lecture transcription. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, pages 217–221, 2012.

[9] T. Kristjansson, A. Culotta, P. Viola, and A. McCallum. Interactive information extraction with constrained conditional random fields. In *Proc. of AAAI*, volume 4, pages 412–418, 2004.

[10] D. D. Lewis and J. Catlett. Heterogeneous uncertainty sampling for supervised learning. In *In Proc. of ICML*, pages 148–156, 1994.

[11] J. R. Lewis. IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use. *International Journal of Human-Computer Interaction*, 7(1):57–78, Jan. 1995.

[12] S. Luz, M. Masoodian, B. Rogers, and C. Deering. Interface design strategies for computer-assisted speech transcription. In *Proc. of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat (OZCHI)*, pages 203–210, 2008.

[13] A. Martínez-Villaronga, M. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing ICASSP 2013*, pages 8450–8454, Vancouver (Canada), 2013.

[14] C. Munteanu, R. Baecker, and G. Penn. Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts. In *Proc. of ACM SIGCHI*, pages 373–382, 2008.

[15] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502, 2006.

[16] H. Nanjo and T. Kawahara. Towards an efficient archive of spontaneous speech: Design of computer-assisted speech transcription system. *The Journal of the Acoustical Society of America*, 120(5):3042–3042, 2006.

[17] J. Nielsen. User interface directions for the web. *Communications of the ACM*, 42(1):65–72, Jan. 1999.

[18] J. Nielsen and J. Levy. Measuring usability preference vs. performance. *Communications of the ACM*, 37(4):66–75, 1994.

[19] Y. Pan, D. Jiang, L. Yao, M. Picheny, and Y. Qin. Effects of automated transcription quality on non-native speakers' comprehension in real-time computer-mediated communication. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1725–1734, 2010.

[20] M. Papadopoulos and E. Pearson. Improving the accessibility of the traditional lecture: an automated tool for supporting transcription. In *Proc. of BCS-HCI*, pages 127–136. British Computer Society, 2012.

[21] G. Riccardi and D. Hakkani-Tur. Active learning: theory and applications to automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 13(4):504–511, 2005.

[22] I. Sanchez-Cortina, N. Serrano, A. Sanchis, and A. Juan. A prototype for interactive speech transcription balancing error and supervision effort. In *Proc. of ACM IUI*, pages 325–326, 2012.

[23] N. Serrano, A. Giménez, J. Civera, A. Sanchis, and A. Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition*, pages 1–13, 2013.

[24] B. Suhm, B. Myers, and A. Waibel. Multimodal error correction for speech user interfaces. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 8(1):60–98, Mar. 2001.

[25] **J. D. Valor Miró**, A. Pérez González de Martos, J. Civera, and A. Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. *Advances in Speech and Language Technologies for Iberian Languages*, pages 237–246, 2012.

[26] F. Wessel, R. Schluter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, 2001.

CHAPTER 5

AUTOMATIC TRANSCRIPTION AND TRANSLATION SYSTEMS FOR MOOCS AND OER

Contents

5.1 Introduction

As discussed in Chapter 1, MOOCs and OER are usually delivered in a single language, and thus imposing a language barrier for those individuals who cannot understand that language. In order to reach a wider audience they should be available in multiple languages, but the full manual translation of the contents is cumbersome. The EMMA project adresses this by using new approaches based in ASR and MT technology.

In fact, the manual transcription of audiovisual materials is a time-consuming and expensive task. A solution is to employ an ASR system to generate an initial draft transcription that could be corrected, reducing so the effort needed for a completely manual approach. As it happens with transcription, MT technology has become a powerful tool to assist lecturers in the translation of textual materials. Also, MOOCs and OER usually incorporate audiovisual content that needs to be transcribed first in order to be translated as subtitles to other languages.

The ASR systems deployed during the EMMA project (see Section 1.3.1), were English, Italian, Spanish, Dutch, Portuguese, French and Estonian, and the MT systems to automatically translate from Italian, Spanish, Dutch, Portuguese, French and Estonian into English and from English into Spanish and Italian. Also, a comparison with mainstream providers like YouTube and Google Translate were performed and reported.

In this chapter, we only report the systems developed in the context of this thesis, which are the Italian and Portuguese ASR systems, and the MT systems Italian-English, English-Italian, Portuguese-English, and Dutch-English. For these languages and language pairs, we describe the linguistic resources that were collected and the training and evaluation process performed to generate these new ASR and MT systems.

5.2 Transcription systems

In this section, Italian and Portuguese ASR systems will be presented respectively. First, we will explain the model estimation during the three periods of the project (P1, P2, and P3), and then, the evaluation and subsequent improvements achieved. For each system we present the resources used, the model estimation process, and the evaluation performed with its results.

5.2.1 Italian

Resources

As in the Spanish ASR system described in Chapter 3, the Italian ASR system just needs annotated audio for the acoustic models and electronic texts for the LM. In this case, first we use annotated Federica Web Learning courses from the Università degli Studi di Napoli Federico II (UNINA), and secondly open AudioBooks of the LiberLiber project. Table 5.1 sums up the basic statistics of acoustic data resources. Also, in order to estimate the LM we

used multiple text data resources freely available. Table 5.2 sums up the basic statistics of these resources.

Table 5.1: Statistics of the annotated speech resource employed in the estimation of the Italian acoustic model.

Corpus	Duration(h)	Words(K)	Vocabulary(K)
Federica	30	400	32
LiberLiber	24	210	10

Table 5.2: Statistics of resources employed to estimate the Italian LM.

Corpus	Sentences(K)	Words(K)	Vocabulary(K)
Federica	10	4000	32
LiberLiber	21	2000	10
EMEA	1081	13000	153
EUbookshop	6490	147000	251
ECB	193	6000	77
OpenSubtitles	15639	97000	133
Europarl	1944	49000	492
Paisa	7555	265000	2933
Wikipedia	17302	285000	5852

Model estimation

In the first period, the acoustic model was generated using TLK [2], and corresponds to the hybrid HMM/DNN. Note that in this case, only a HMM/DNN trained from the CMLLR data is estimated. The characteristics of the P1 Italian acoustic models are the following:

- Standard model (1-pass)
  - 5077 tiedphoneme 3-state HMM/GMM with a 128-mixture component Gaussian per state.
- CMLLR models (2-pass)
  - 5063 tiedphoneme 3-state HMM/GMM with a 128-mixture component Gaussian per state.
  - A 5-layer HMM/DNN with the following architecture: 48 input cells, 5063 output cells and 3000 cells in each internal layer.

In the second period, the Italian ASR system was improved both, the acoustic and language model. First, the standard acoustic model has been changed by a HMM/DNN model. Moreover, the LM has been improved using a RNNLM [5] estimated on the Federica dataset, which is content related with UNINA MOOCs. Last, an error analysis on the automatic transcriptions showed that silence/speech detection was not being properly performed, leaving some parts of the speech untranscribed. This problem was solved in this period by applying an automatic segmentation process to improve the silence/speech detection.

The Italian ASR system was further improved in the third period using the mDNN training using also the Spanish speech data. The Spanish ASR corpus consists in 97 hours that was added to the 54 hours of the Italian speech corpus for a total of 151 hours of speech training data. Moreover, we improved our mDNN using 8 iterations of the sequence-discriminative training on the Italian corpus.

Note that the lexical model was obtained using a phonetic transliterator based on the IPA source phonetic rules [3], that was built as part of the EMMA project. More details about all of these techniques can be found in Section 2.3.

Evaluation

From the provided 30 hours of annotated speech data provided by UNINA, 8 hours were devoted to the development and test purposes. Although the Federica dataset is domain-related to the videos recorded for UNINA MOOCs courses, specific annotated videos from UNINA MOOCs are preferable. These videos were automatically transcribed using the Italian ASR system and intended to be post-edited in the transLectures player. However, these new videos provided by UNINA did not possess enough sound quality (low gain and echo) to generate acceptable transcriptions. Table 5.3 depicts the main statistics of the Federica evaluation set.

Table 5.3: Basic statistics of the evaluation dataset for the Italian ASR system.

Set	Videos	Duration	Words (K)	Vocabulary (K)
Development	3	4h	35.6	4.9
Test	3	4h	33.4	5.0

In the first period (P1), it should be noticed that only the first two passes of the recognition process were applied. In the second period (P2) we added the third recognition step on which the network were adapted to the speaker. Also, the recognition process was improved by re-scoring transcription hypotheses with the RNNLM. Finally, an automatic segmentation process was employed in order to improve the silence and phrase detection in the test set. In the third period (P3), the new Italian ASR system was evaluated on the test set.

In order to assess these improvements, an empirical evaluation was performed using the Federica dataset. Table 5.4 shows the WER results obtained on the test dataset over the three periods.

Table 5.4: Results of the conventional ASR system for Italian on the test set.

Set	P1	P2	P3
Test	21.2	17.7	17.1

As observed in Table 5.4, the transcription quality of the Italian ASR system is lower than 20 points of WER, which can be considered as a high-quality state-of-the-art ASR system. The WER of the YouTube ASR system on the same test is 31.6, which is almost twice the WER achieved by our conventional system. This corroborates that our ASR system is in range with those at the state-of-the-art.

As with Spanish system, an efficient version was generated in order to improve the response time of the transcription process. As described in Section 2.3.3, a pruned version of the recognition process, which only performs two recognition passes, is employed together with a RNNLM re-scoring to reduce the time cost while only slightly increasing the WER. Table 5.5 shows the WER and RTF results obtained on the test dataset.

Table 5.5: Results in quality and efficiency for the Italian ASR system on the test set.

System	WER	RTF
P2 conventional	17.7	4.8
P2 efficient	18.6	0.8
P3 conventional	17.1	4.8
P3 efficient	17.3	1.4

As expected, results are slightly worse compared with our best system, but the speed has been greatly improved. It is important to note that small WER difference is not noticed by users, so the quality in terms of human evaluation, would remain the same as the best system, while the response time is more than three times faster.

5.2.2 Portuguese

Resources

Multiple resources have been employed to estimate the Portuguese acoustic model. Table 5.6 sums up the basic statistics of all these resources. Statistics for the textual resources devoted to the estimation of the Portuguese LM are provided in Table 5.7. It must be noted that all these text databases are freely available.

Model estimation

As for Spanish and Italian, the Portuguese ASR system required the training of acoustic and language models. In the first period only ELDA, LAPS-UFPA and UAB corpus were employed

Table 5.6: Statistics of the resources employed in the estimation of the Portuguese acoustic model.

Corpus	Duration(h)	Words(K)	Vocabulary(K)
ELDA	17	163	19
LAPS-UFPA	8	77	7
UAB	2.5	44	6
TEDx	30	439.3	17.5

Table 5.7: Statistics of resources employed to estimate the Portuguese LM.

Corpus	Sentences(K)	Words(K)	Vocabulary(K)
ELDA	17	163	19
LAPS-UFPA	9	77	7
UAB	2	40	6
EMEA	996	12602	42
EUbookshop	4157	99535	546
ECB	200	5437	43
OpenSubtitles	20493	116422	431
Europarl	1983	50085	149
MsgImooc	7464	100112	505
Ceten	1645	26280	217
Wikipedia	10823	168121	1551

to train the acoustic model. This model corresponds to a hybrid HMM/DNN model described in Section 2.3.1. Similarly to the Italian ASR system, the characteristics of the first period Portuguese acoustic models are the following:

- Standard model (1-pass)
  - 2964 tiedphoneme 3-state HMM/GMM with a 128-mixture component Gaussian per state.
- CMLLR models (2-pass)
  - 3020 tiedphoneme 3-state HMM/GMM with a 128-mixture component Gaussian per state.
  - A 5-layer HMM/DNN with the following architecture: 48 input cells, 3020 output cells and 3000 cells in each internal layer.

The language model for Portuguese corresponds to a linear mixture of interpolated 4-gram models (see Section 2.2).

In the second period the Portuguese ASR system was improved by refining the acoustic model. First, 30 hours of annotated speech were added for the estimation of the acoustic model. These new annotated samples correspond to TEDx talks (see Table 5.6). As a result, the acoustic model is now trained with 50 hours of speech.

In addition, based on the experiments performed in the first period, it was found out that the baseline ASR system, which performs three passes in recognition phase, obtained worse results than a simpler one-pass system. For a more detailed description about the multiple pass automatic transcription process please refer to Section 2.3.3. This is mainly due to the presence of background music, the different variants of spoken Portuguese present in the training corpus, and the multispeaker nature of these videos. All these specific features generate higher uncertainty in the system. In fact, the second and third recognition passes that deal with speaker adaptation, when there is no sufficient data to reliably estimate each speaker, degrade the performance of the system.

In the third period, in collaboration with the UAB, the Portuguese variant (European, Brazilian or other) of 35 hours from the training set was annotated. These variants are very different from each other, which leads to a serious degradation in the system performance when they are not processed separately [6] [1]. Therefore, 6.2 hours of audio samples of the European Portuguese variant were devoted to train a new acoustic model. However, the results were not improved due to the lack of resources, so we reverted to the mixed-variant model. Then, we realigned all the training data in order to obtain more accurate acoustic models, using the same training process as we performed in the second period.

Note that phonetic transcriptions of the annotated resources were obtained using a phonetic transliterator based on the IPA source phonetic rules built as part of the EMMA project.

Evaluation

At the beginning of the project, the Universidade Aberta (UAB) provided two hours of transcribed videos. It is important to note that these videos contain background music along with speech, making more difficult to use this annotated data. These two hours of speech were devoted to the development and test sets for evaluation purposes.

From the limited speech data resources for Portuguese, an ASR system was trained to transcribe 6 videos involved in UAB EMMA MOOC course *Climate Change*. However, the automatic transcriptions were not worth to be post-edited and had to be transcribed from scratch because of the low transcription quality. The reason for these low-quality transcriptions is that loud background music in these videos made not possible to generate acceptable transcriptions. This set of 6 videos constitutes the EMMA Test set. Basic statistics of the evaluation datasets can be observed in Table 5.8.

The ASR system for Portuguese was evaluated on the evaluation sets described above. However, in order to isolate how the background music affects the Portuguese ASR performance, a fully-manual segmentation and speech/non-speech annotation for each segment were carried on the evaluation datasets. Non-speech segments were discarded to compute WER

Table 5.8: Basic statistics of the evaluation dataset for Portuguese ASR

Set	Videos	Duration	Words (K)	Vocabulary (K)
Development	2	1h	8.0	1.8
Test	2	1h	7.5	1.7
EMMA Test	6	18m	2.4	0.9

figures. All in all, speech segments still contain background music that produce lower quality transcriptions.

In the first period, as observed in Table 5.9, the overall quality of the system is low. The reason for these low-quality transcriptions is that loud background music in these videos made not possible to generate acceptable transcriptions. WER is even worse for the EMMA test set in which music was louder than speech.

In the second period, an improvement on quality was obtained by adding LibriVox and TED datasets in the training, and skipping adaptation steps, which was degrading the performance of the ASR system.

Table 5.9: Results of the ASR system for Portuguese on the evaluation datasets.

System	WER Test	WER EMMA
P1	45.9	67.3
P2	43.0	63.1
P3	42.0	62.4

In the third period, a European-Portuguese ASR system trained on 6.2 hours was developed, but there was not enough speech data to properly estimate a variant-specific acoustic model, and the performance of the variant-mixture acoustic model still was better. Therefore, we reverted to P2 version; and providing better alignments for the speech corpora we slightly improved the results, as can be observed in Table 5.9.

In order to get a better idea of the quality of our system, we performed a comparison with YouTube in a similar setting as in other languages. The test set of this task was recognised with a 62.3 WER points in terms of quality. This shows that after all, our ASR system performs better than other state-of-the-art systems like YouTube.

5.3 Translation systems

In this section we present the translation systems from Italian, Portuguese and Dutch into English and English into Italian. First, we will explain the target language model estimation, and then, all the MT systems will be described and evaluated in terms of BLEU (see Section 2.5 for more details). Note that in most systems, as indicated in each of them, we used the human

BLEU to evaluate them. This BLEU is that obtained from the review by a human of the automatic translation of the test set, which usually performs the minimum editions possible.

Note that the LM for the translation systems is that of the target language. In our case, this means that apart from the Italian LM trained as explained in Section 5.2, we need to train an English language model to be included in the translation systems. All these specific models were also interpolated with a LM trained on the target part of the selected parallel data.

5.3.1 English language model

The English LM described in this section has been used for all SMT systems in which their target language was English, specifically in Portuguese-English, Dutch-English and Italian-English systems; while for English-Italian we used the LMs described in in Section 5.2 as explained before.

Table 5.10 depicts the basic statistics of all text databases employed in the estimation of English model. Note that, these figures have been computed after preprocessing the raw text to reduce the complexity.

Table 5.10: Statistics of resources employed to estimate the English language model

Corpus	Sentences(K)	Words(M)	Vocabulary(K)
PHP	32	0.1	4.3
Tatoeba	18	0.1	7.6
EUconst	10	0.2	6.5
EUTV	158.3	1.5	25.0
DPC	176.7	3.0	61.2
ECB	126.5	3.1	28.2
TED	379.2	3.1	51.8
EMEA	1090.9	13.8	44.0
JRC	1240.2	32.0	542.2
Europarl	2026.1	55.0	137.0
DGT	4900.2	85.5	1468.2
EUbookshop	5964.6	146.9	853.3
OpenSubtitles	16998.8	129.5	359.2
Wikipedia	82617.6	1470.3	6148.7
Google	-	458200	10300

All these resources were employed to estimate the LM following the process described in Section 2.2. However, instead of a 4-gram LM, a 5-gram was estimated, as a longer word history has been shown to yield better translation results. This 5-gram LM is linearly interpolated with a LM trained on the English part of the selected parallel data and optimised on held-out monolingual in-domain data.

5.3.2 Italian-English

Resources

As said in Section 2.4, nowadays there are large freely available Italian-English parallel texts. Table 5.11 describes the basic statistics of all parallel resources employed in the translation model estimation. The FedericaParallel corpus is an in-domain corpus that UNINA provided for the training of the Italian-English system.

Table 5.11: Statistics of resources employed to estimate the Italian-English translation model

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		It	En	It	En
FedericaParallel	2.8	0.05	0.05	7.1	5.3
ECB	193.1	5.8	5.4	77.6	62.1
EMEA	1081.1	13.4	12.1	153.5	130.1
EUbookshop	6490.0	147.3	144.6	2513.7	2329.9
Europarl	1944.8	48.9	50.7	492.3	380.0
OpenSubtitles	15639.8	97.2	104.8	1331.1	1013.0
DGT	4900.2	88.0	97.8	1458.5	1467.5
EUconst	10.0	0.1	0.1	13.9	11.8

Model estimation

In the first period, as explained in Section 2.4, we selected from the out-domain resources some domain-related data to build a parallel in-domain set. In fact, bilingual selection techniques described in Section 2.4 were employed to select again relevant data from out-domain resources, obtaining a better dataset. This data selected, as well as the previously generated in-domain set, were used as the training data for the translation and language models. The training of the MT system was performed using the Moses toolkit [4].

In the second period, the MT system was improved by extracting a new parallel text training set using the approach described in Section 2.4.1. Concretely, we used the INS, followed by a bilingual Axelrod selection to increase the amount of related text-data selected. After this selection process was applied, a parallel text dataset was obtained, and then, it was employed to estimate the MT system.

In the third period, this process was automatised. The statistics of the final training set are presented in Table 5.12.

Table 5.12: Statistics of selected data employed to estimate the Italian-English translation model

Sentences(M)	Words(M)		Vocabulary(K)	
	It	En	It	En
6.6	140.8	141.6	921.3	852.8

Evaluation

In the first period, some of the UNINA reviewed translations from MOOC video transcriptions were devoted to the test set, while a part of previously available in-domain parallel texts (Federica parallel corpus) were included in the development set. We compared the MT systems generated using two different selection techniques: Bilingual Axelrod and bilingual Moore.

In the second period we added a new second test set with the automatic translated and post-edited texts related to UNINA courses. These texts were first automatically translated with our system, and then reviewed by the UNINA lecturers and staff. Table 5.13 shows basic statistics of the Italian-English evaluation datasets.

In the third period the system is basically the same, but the translation process is fully automated from the training data.

Table 5.13: Basic statistics of annotated data for the Italian-English SMT evaluation.

Set	Sentences	Words (K)		Vocabulary (K)	
		It	En	It	En
Development	1181	22.4	23.4	3.9	3.2
Video Test	446	7.4	7.8	1.5	1.3
Document test	1028	20.4	20.7	3.3	2.7

The Italian-English MT system parameters were tuned on the development, and these MT systems translated the test set in its entirety in order to compute BLEU scores on the resulting automatic translations. Table 5.14 shows BLEU scores obtained on the evaluation datasets.

Table 5.14: BLEU scores of the Italian-English SMT systems.

Selection technique	Development	Video Test	Documents Test
P1 - Bilingual Moore	15.8	40.1	-
P1 - Bilingual Axelrod	18.6	42.9	-
P2 - INS+Axelrod	-	47.8	40.7
P3 - Auto INS+Axelrod	-	48.0	40.7

As observed in Table 5.14, the results were greatly improved for this language pair by using the intelligent data selection approach. This led to the generation of better translations, and consequently to user-effort reduction in the subsequent review process.

Finally, we compared the quality of our system with that obtained by Google Translate, in order to obtain a direct comparison with a state-of-the-art system. Google Translate obtained a 37.5 of BLEU on the Video Test set and 35.5 in the Documents Test, which is far worse than that of our system .This difference is explained by the application of domain adaptation techniques.

5.3.3 Portuguese-English

Resources

To estimate the MT system for the Portuguese-English language pair, we used parallel resources. Table 5.16 describes the basic statistic of all out-domain parallel resources employed in the Portuguese-English translation model estimation. It must be noted that no in-domain parallel dataset was provided.

Table 5.15: Statistics of resources employed to estimate the Portuguese-English translation model

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		Pt	En	Pt	En
AMARA	230.1	2.3	2.3	113.9	85.7
ECB	202.0	6.2	5.7	81.9	62.7
EMEA	1082.2	13.6	12.1	128.8	130.3
EUbookshop	4172.2	102.5	96.4	1853.2	1677.8
Europarl	2001.6	50.9	50.3	474.7	349.7
DGT	4900.2	91.2	97.8	1440.7	1467.4
JRC	1236.8	33.7	31.9	588.8	538.3
OpenSubtitles	20508.9	120.1	132.9	1474.0	1190.1

Model estimation

In this translation pair we introduce the problem of creating an in-domain parallel data, that were not provided to us. We adressed this problem using the intelligent data selection approach described at Section 2.4.1. Specifically, in the first period an initial INS selection, in which no parallel data is necessary, was performed. This initial selection helps us to create a larger in-domain dataset so the Axelrod selection technique can be applied using the available monolingual in-domain data.

Then, in the second period, the data selected from the out-domain was used as in-domain, and a new selection from the remaining out-domain corpora was performed using the Axelrod technique. The selected out-domain data together with the initially generated in-domain data was used as the training data for the translation and language models. This approach, that were automatised in the third period, is the same used in other translation pairs. The statistics of this dataset can be found in Table 5.16.

Table 5.16: Statistics of selection data employed to estimate the Portuguese-English translation model

Sentences(M)	Words(M)		Vocabulary(K)	
	Pt	En	Pt	En
6.6	84.0	89.5	621.2	518.5

As in previous translation systems, the translation model was trained using the Moses toolkit.

Evaluation

Automatic translations of UAB MOOCs videos became a test set used for evaluation purposes. The basic statistics of the test sets for the Portuguese-English MT evaluation are presented in Table 5.17.

Table 5.17: Basic statistics of annotated data for the Portuguese-English MT evaluation.

Set	Sentences	Words (K)		Vocabulary (K)	
		Pt	En	Pt	En
Test	83	2.7	2.5	0.9	0.8

The same evaluation procedure applied in previous language pairs was followed for the Portuguese-English language pair. BLEU scores on the test sets are reported in Table 5.18.

Table 5.18: BLEU scores of the Portuguese-English MT systems.

System	Test
P1	52.8
P2	55.5
P3	51.7

As observed in Table 5.18, BLEU scores reflect that the performance of Portuguese-English MT system is excellent.

Note that the automatic intelligent data selection approach presents a lower BLEU score. One of the reasons for this performance degradation in the automatic adaptation process is explained in the Section 5.3.5 for the English-Italian translation pair. In addition, the test set defined for EMMA only contains 83 sentences when the usual number of sentences in a test set is about one thousand. Small test sets are very sensitive to modifications in the system since few errors are magnified by the size of the test set.

For the sake of comparison and as in other languages, we performed a quality comparison with Google Translate. The result of Google Translate when translating the EMMA test is 45.4 of BLEU, which is worse than our system. Again, the adaptation to the content of the UAB courses significantly improves the accuracy of our translation system.

5.3.4 Dutch-English

Resources

In the third period of EMMA we developed from scratch the Dutch-English translation system to translate MOOC contents of the Open University in the Netherlands (OUNL). To this purpose, OUNL provided a significant amount of Dutch resources corresponding to web content and bibliography related to their MOOC courses. The main drawback of these resources is that they are monolingual, i.e. only Dutch transcriptions were available. Nevertheless, a large amount of general out-domain resources were collected. Table 5.19 describes the basic statistics of in-domain (first two rows) and out-domain (third and subsequent rows) parallel resources employed in the translation model estimation.

Model estimation

As in previous language pairs, in Dutch we have limited in-domain resources. For this reason, we develop the MT system following the same procedure as we used in previous pairs, to create an in-domain corpus from the out-domain corpora available, using INS and Axelrod selection techniques. The estimation of the MT system was performed using the Moses toolkit, and the LM was interpolated in a similar fashion to other translation pairs to improve the final system performance.

Evaluation

First of all, with an initial general-purpose MT system we assisted the OUNL revision team in the generation of an in-domain parallel dataset by post-editing the automatic translations. With this procedure, we achieve reference correct translations using the transLectures player. This set of reference translations was used as a test set, and its basic statistics can be seen in Table 5.20.

Table 5.19: Statistics of resources employed to estimate the Dutch-English translation model

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		Nl	En	Nl	En
OUNL	306.6	1.5	-	51.7	-
DomainCorpusNLDocs	304.0	1.4	-	49.0	-
PHP	32	0.1	0.1	5.5	4.4.
Tatoeba	18	0.1	0.1	9.7	7.6
EUconst	10	0.2	0.2	7.9	6.5
EUTV	158.3	1.7	1.5	38.0	25.0
DPC	176.7	3.1	3.0	106.1	61.2
ECB	126.5	3.3	3.1	53.7	28.2
TED	379.2	3.5	3.1	88.6	51.7
EMEA	1090.9	12.3	13.8	73.8	43.9
JRC	1240.2	32.7	32.0	253.6	168.7
Europarl	2026.1	53.8	55.0	358.5	136.6
DGT	4900.2	97.8	85.5	697.4	557.6
EUbookshop	5964.6	150.1	146.9	1095.2	845.3
OpenSubtitles	16998.8	147.3	129.5	534.0	523.992

Table 5.20: Basic statistics of the updated test set for the Dutch-English SMT evaluation.

Set	Sentences	Words (K)		Vocabulary (K)	
		Nl	En	Nl	En
Test	4048	46.0	46.5	5.1	3.9

Table 5.21 shows BLEU scores on the test set for the Dutch-English MT system described. As we can observe, the quality of our MT system is good enough to generate quality translations. Also, we performed a comparison with Google Translate, which obtains 33.4 BLEU on the same test set. Therefore, we can consider our Dutch-English MT system at the level of state-of-the-art.

Table 5.21: BLEU scores of the Dutch-English MT system.

System	Test
P3	43.5

5.3.5 English-Italian

Resources

This language pair was introduced in EMMA at the second period, and it was relatively simple to generate an MT system for it, as it can be obtained using the same data as that of the Italian-English MT system. Table 5.22 describes basic statistics of all parallel resources employed in the translation model estimation.

Table 5.22: Statistics of resources employed to estimate the English-Italian translation model

Corpus	Sentences(K)	Words(M)		Vocabulary(K)	
		It	En	It	En
FedericaParallel	2.8	0.05	0.05	7.1	5.3
ECB	193.1	5.8	5.4	77.6	62.1
EMEA	1081.1	13.4	12.1	153.5	130.1
EUbookshop	6490.0	147.3	144.6	2513.7	2329.9
Europarl	1944.8	48.9	50.7	492.3	380.0
OpenSubtitles	15639.8	97.2	104.8	1331.1	1013.0
DGT	4900.2	88.0	97.8	1458.5	1467.5
EUconst	10.0	0.1	0.1	13.9	11.8

Model estimation

In the second period, in order to train the MT system we used the same data selection approach as in the Italian to English system, that is an INS, followed by a Bilingual Axelrod selection. After selection being applied a parallel text dataset is obtained and used to train the MT system. The statistics of the selection obtained are presented in Table 5.23.

Table 5.23: Statistics of selection data employed to estimate the English-Italian translation model

Sentences(M)	Words(M)		Vocabulary(K)	
	En	It	En	It
6.5	134.5	133.0	808.1	871.7

The training of the MT system was performed using the Moses toolkit. Last, a second LM generated using Italian monolingual data was added. This model correspond to the LM employed in the Italian ASR system.

In the third period, as in the case of Italian into English, this process was automatised.

Evaluation

This section describes the evaluation of the MT system developed for automatically translating English text into Italian. The English-Italian evaluation set was created from the review process of the UPV MOOCs: *Search on Internet* and *Excel 2010*. Table 5.24 shows basic statistics of the parallel text for the English-Italian MT evaluation.

Table 5.24: Basic statistics of the updated test set for the English-Italian SMT evaluation.

Set	Sentences	Words (K)		Vocabulary (K)	
		En	It	En	It
Test	5823	60.8	54.0	3.9	5.6

As observed in Table 5.25, the scores obtained are high. However, the result of the automatic intelligent data selection approach (P3) is slightly worse than the manually intelligent data selection (P2). The reason behind this difference is the order in which word alignments are computed in the training process. In the manually generated system, alignments are estimated on the selected training set from the out-domain corpora, while in the automatic system for efficiency reasons alignments are estimated on the complete out-domain corpora and then, the selection is applied. We believe that the alignments computed on the complete out-domain corpora contain more noise than those computed on the selected training set, since incorrect sentence pairs were not filtered out by the selection process. All in all, the benefits derived from the capability of automatically generating customised course-adapted MT system outweighs this minor decrease in translation quality.

Table 5.25: BLEU scores of the English-Italian MT system.

System	Test
P2	52.6
P3	51.3

As usual, we compared our MT system with that provided by Google Translate, which obtained 44.1 of BLEU. With this data we can confirm that our system is in the state-of-the-art.

5.4 Conclusions

In this chapter we have presented all the resources and techniques employed to build ASR and MT systems for the Italian and Portuguese ASR systems, and the MT systems Italian-English, English-Italian, Portuguese-English, and Dutch-English.

The Italian ASR system achieved excellent results thanks to technology improvements and adaptation. However, the Portuguese ASR system was poorly trained because little annotated speech data was available and in addition, the presence of background music made very difficult proper ASR. MT systems have at their disposal large amounts of out-domain data for all the language pairs, thanks to the existence of huge parallel corpus freely available on the Internet. However, in-domain parallel data is in most cases scarce and little, if any, educational parallel texts are available. To solve this problem, intelligent selection techniques are applied to create or enlarge the amount of in-domain parallel data.

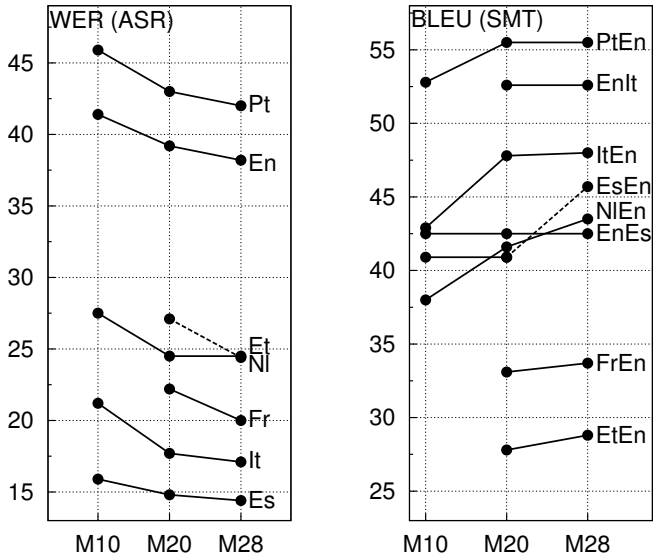

Figure 5.1: Progress for all languages in ASR on the left, given in terms of WER (the lower, the better) and in MT on the right, in terms of BLEU (the higher, the better).

In Figure 5.1 we represent the progress of the ASR and MT systems developed during the EMMA project in terms of WER and BLEU, respectively, on the test set for each task. For the sake of information, the data of this figure is not limited to the systems trained as part of this thesis, but to all involved during the project. This plot gives us an idea on how the systems have been improved and which the state of each language is when compared to one another. Most of the ASR systems are below 30% of WER, which can be considered state-of-the-art results for this task. Almost all MT systems are over 35 points of BLEU, which can be considered a good result and in range with commercial MT services. In any case, a more detailed assessment in quality and review time with real-life users will be carried out in Chapter 6, as part of the work of this thesis.

Bibliography

[1] A. Abad, H. Meinedo, I. Trancoso, and J. Neto. Transcription of multi-variety portuguese media contents. In *Computational Processing of the Portuguese Language*, pages 409–420. Springer, 2012.

[2] M. A. del Agua, A. Giménez, N. Serrano, J. Andrés-Ferrer, J. Civera, A. Sanchis, and A. Juan. The transLectures-UPV toolkit. In *Proc. of IberSpeech 2014*, Las Palmas de Gran Canaria (Spain), 2014.

[3] IPA Source. Diction Help of the IPA Source. `http://www.ipasource.com/diction_help`, 2014.

[4] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. Moses: Open source toolkit for statistical machine translation. In *Proc. of ACL*, Prague (Czech Republic), 2007.

[5] T. Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, Brno Univ. of Technology (FIT), 2012.

[6] J.-L. Rouas, I. Trancoso, C. Viana, and M. Abreu. Language and variety verification on broadcast news for portuguese. *Speech Communication*, 50(11):965–979, 2008.

CHAPTER 6

EVALUATION ON THE REVIEW OF MULTILINGUAL VIDEOS FOR MOOCS AND OER

Contents

6.1 Introduction

In Chapter 5 we presented advanced ASR and MT systems that has been used to produce multilingual videos for MOOCs and OER. Obviously, this task requires expertise, resources and tools from ASR and MT, but also additional components and experience for their proper integration into real-life educational environments. This chapter provides a description on the integration of the resources and tools that we have used and, more importantly, a comprehensive evaluation of the results achieved in two real-life case studies, a MOOC platform and a large video lecture repository (see Section 1.3), as well as the impact that multilingual videos have had in these case studies [20]. Note that the comparisons performed here are done with more videos than those performed in the previous chapter, which was limited to test sets.

This chapter is organised as follows. First, the integration of our technology is summarised in Section 6.2. Then, detailed results on transcription and translation quality are provided in Section 6.3, also including comparative results with mainstream providers, which in case of Italian and Dutch are similar to those presented in the previous chapter. In Section 6.4, these results are followed by a thorough evaluation of transcription (translation) reviewing time for each language (language pair) considered separately, and also across all languages considered. Section 6.5 is devoted to the impact these systems, tools and integration components have had in the case studies. Finally, the main conclusions drawn are summarised in Section 6.6.

6.2 Integration of ASR and MT systems

As explained in Chapter 5, general-purpose ASR and MT systems (e.g. YouTube and Google Translate) can achieve reasonable results in many cases, but model adaptation often leads to much more accurate results [10]. Speaker and topic adaptation are conventional approaches to the adaptation of, respectively, acoustic and language models [9, 12]. Translation models, on the other hand, are often adapted by mining (parallel) sentences from large out-of-domain corpora which somehow resemble sentences in related in-domain corpora [5].

Integration of adapted ASR and MT systems into a video lecture repository is not straightforward. To this end, a free, open-source solution called the transLectures-UPV Platform (TLP) can be applied [1]. Generally speaking, TLP is a middleware software making transcription and translation services easily available to MOOCs and OER. It comprises four main components: a PHP/HTML5 media player/editor for reviewing subtitles; a web service to integrate TLP services into media repositories; an ingest service including core functions to manage multilingual media; and TLP database to support the web and ingest services.

Figure 6.1 shows three TLP use cases for a video lecture repository in which the role of each TLP component is clearly exemplified. The first use case (left) consists in adding a new recording. In this case, the video lecture repository uploads the new media to the TLP server using the *ingest* interface of the web service. Then, the ingest service runs the appropriate ASR/MT systems and stores both the new media and its subtitles in the TLP database.

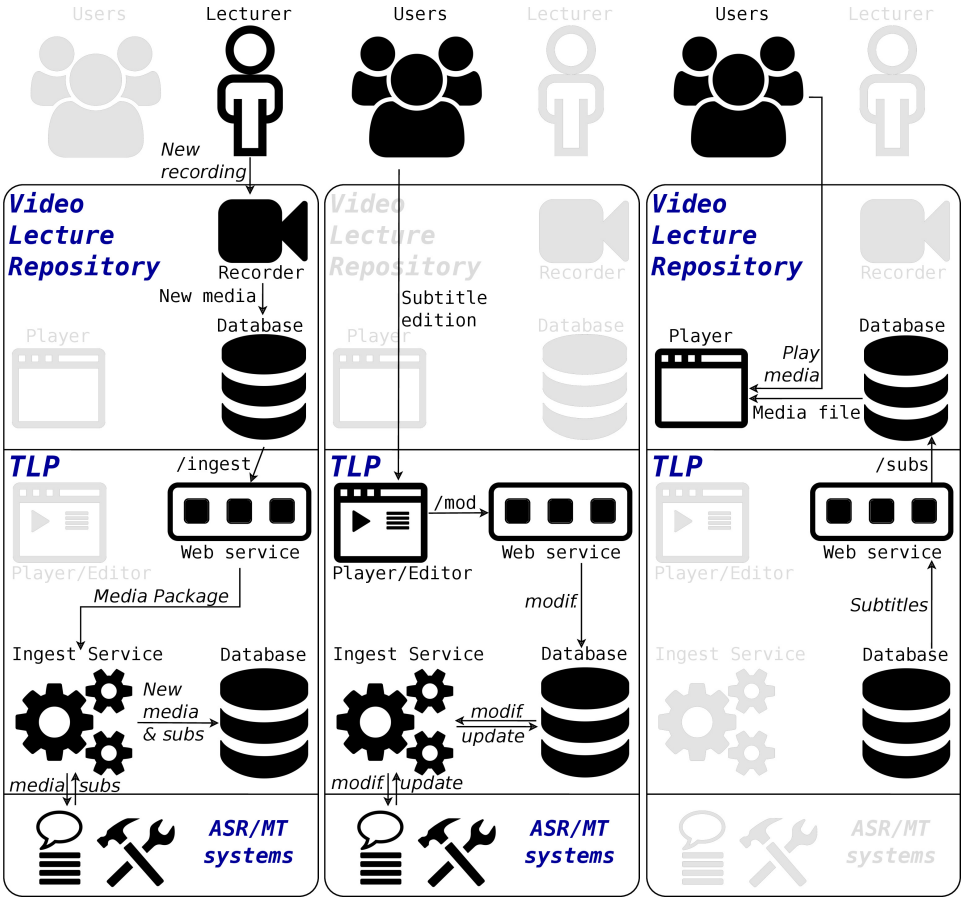

Figure 6.1: Three TLP use cases: adding a new recording (left), reviewing subtitles (centre), and retrieving multilingual subtitles (right).

In the second use case shown in Figure 6.1 (centre), video lecture subtitles are manually reviewed by using the TLP player/editor. All revisions, either modifications or simple validations, are added to the TLP database through the *mod* interface of the web service. Also, they are processed by the ingest service to improve ASR/MT models and update all non-reviewed subtitles. The third and final use case shown in Figure 6.1 (right) is on retrieving multilingual subtitles. In this case, the video lecture repository retrieves the relevant multilingual subtitles from the TLP database via the *subs* interface of the web service.

Figure 6.2 shows two examples of use of the multilingual TLP player/editor with its default layout. The first example (top) is an editor of transcriptions. The video and its segmentation are displayed on the left, while transcriptions are shown on the right, with each individual transcription segment in a separate text edit box. Segments can be inserted, deleted or merged

either on the left or right. Their time synchronisation can be adjusted on the segmentation on the bottom left, while the actual transcriptions can be modified in the edit boxes on the right. The second example (bottom) is an editor of transcriptions and translations. It is analogous to the first example, but with translations also available on the right. Apart from the two use modes illustrated in Figure 6.2, the multilingual TLP player/editor can also be used as an editor of translated text contents.

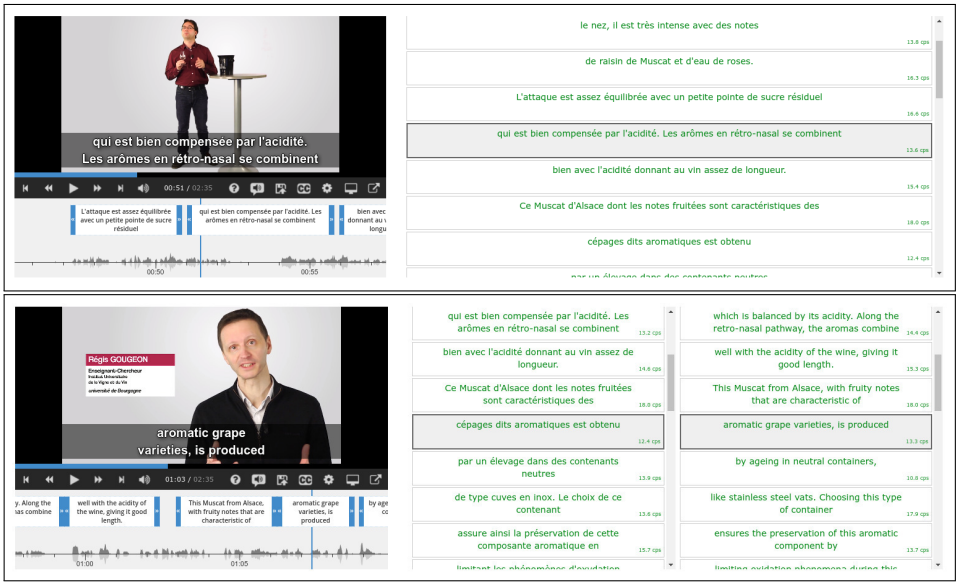

Figure 6.2: Multilingual editor of transcriptions (top) and translations (bottom).

The above tools and integration components can be of great help to generate multilingual videos in cases similar to those described above. However, it is not trivial to develop new systems for new cases and, of course, it is much more difficult to obtain accurate adapted systems with limited experience. To help potential users in adopting these tools, the systems developed for the EMMA platform and the UPV media repository can be freely tried through the so-called Transcription and Translation Platform (TTP) [3] (see Figure 6.3).

Broadly speaking, TTP is an online platform for automated and assisted multilingual media processing, and particularly for subtitling and text translation. Here it can be seen just an example of TLP-based integration for the generation of multilingual videos.

6.3 Transcription and translation quality

In this section, we assess the quality of automatic transcriptions and translations ASR/MT systems for videos originally in 5 languages drawn from the UPV media repository and the

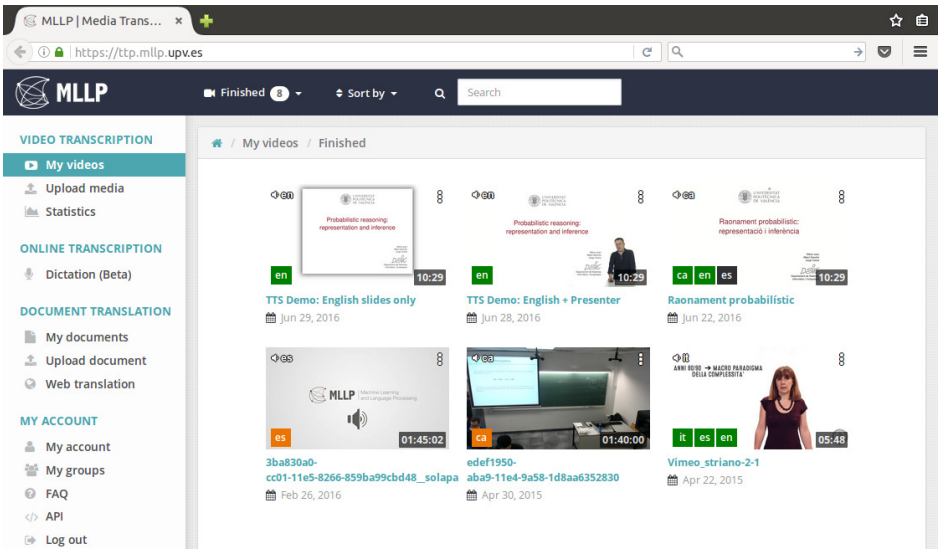

Figure 6.3: Screenshot of the TTP.

EMMA platform: Spanish, Italian, Dutch, French, English. Results are not available for Portuguese because the review of transcriptions and translations was not performed on TTP.

As mentioned in Chapter 5, videos are included in MOOCs from the EMMA platform provided by four European universities: UNINA, OUNL, Université de Bourgogne and UPV. In addition, a comparative evaluation of transcription and translation quality with mainstream providers of ASR and MT technology, i.e. YouTube and Google Translate, is also presented.

6.3.1 Transcription quality

Transcription quality was measured with the widely accepted WER criterion explained in Section 2.5. In this regard, it must be noted that achieving error-free transcriptions is totally unrealistic, even if they are manually produced. On the contrary, it is more realistic to expect a WER of about 10% from commercial, manual transcription services [11]. From a practical point of view, automatic transcriptions of WER equal or less than 25% convey enough correct information to be useful [13], and professional stenographers prefer them to manually transcribe from scratch [4].

For this reason, an experiment was conducted to investigate the usability of the ASR-based transcription in real applications in Spanish, English, Italian, Dutch and French. The Spanish videos were reviewed by the lectures involved in the DeX action plan and the English videos were reviewed by volunteers. Also, both come from the UPV media repository. Italian, Dutch and French videos were included in MOOCs delivered on the EMMA platform, and were

transcribed using TTP by the institutions that offers the courses. Table 6.1 shows the number of videos, duration (in hours) and WER (± standard deviation) for each transcribed language.

There is a significant number of Spanish videos since more than 90% of the videos in the UPV media repository are in Spanish. In addition, as mentioned before, the review of transcriptions (and translations) is incentivised by the DeX action plan.

The average duration of videos for all languages except for Dutch is less than 10 minutes. Dutch videos last more than 35 minutes on average and the format of the video presentation is different from that in the other languages. Dutch videos are interviews with usually two speakers sitting around a table, while in the other videos a single speaker stands in front of the camera.

Table 6.1: Videos, duration (hrs.) and WER (± std. dev.) for each language.

Language	Videos	Hrs.	WER
Spanish	207	24.7	18.4 ± 6.4
Italian	13	1.2	25.7 ± 6.4
English	25	3.5	21.9 ± 8.5
Dutch	11	6.9	29.4 ± 9.2
French	21	2.1	23.2 ± 8.3

From the results in Table 6.1, we can observe that the quality of Spanish transcriptions is the highest, followed by English and French, all three being below 25%. Italian is just above 25% of WER and Dutch is the highest WER figure, but below 30% of WER. In the case of Dutch, we believe that the higher WER figure is explained by the presence of more than one speaker in the videos, which harms the acoustic adaptation to the speaker, not being so effective as in the rest of the videos in which a single speaker appears.

6.3.2 Translation quality

As in transcription, translation quality is often measured with an error criterion: the so-called TER [17]. More details of this metric can be found in Section 2.5. As with the WER, it must be noted that achieving error-free translations, either automatic or manual, is unrealistic. In addition, in the case of MT it is generally accepted that source sentences can be manually translated into many different yet correct ways, and thus a correct translation for a certain reviewer might no be the preferred (correct) translation for another. As the TER is computed from only one correct reference, it is considered a pessimistic criterion. From a practical point of view, automatic translations with TER figures below 50% are worth post-editing, instead of translating from scratch [18, 19].

Table 6.2 shows the number of videos, duration (in hours) and TER (± standard deviation) for each translation pair. All videos were automatically translated and then reviewed. The Spanish videos are part of the UPV media repository and were reviewed by lecturers involved

in the DeX action plan. The English videos translated into Spanish are from two EMMA MOOCs originally in Italian, then translated into English, and now for this work translated into Spanish. Conversely, the English videos translated into Italian are from two EMMA MOOCs originally in Spanish, then translated into English, and finally translated into Italian. Note that, as mentioned before, there are (four) MOOCs available in three languages (Italian, English and Spanish). Finally, the Dutch and French videos are also from EMMA MOOCs translated into English.

Table 6.2: Videos, duration (hrs.) and TER (± std. dev.) for each translation pair.

Tl. pair	Videos	Hrs.	TER
Es → En	101	10.8	33.2 ± 14.4
En → Es	29	2.5	27.0 ± 19.9
It → En	14	1.6	37.5 ± 8.2
En → It	121	6.5	33.8 ± 8.0
Nl → En	5	3.5	30.7 ± 13.4
Fr → En	8	0.9	58.9 ± 5.2

From the results in Table 6.2, it is clear that, apart from the French→English pair, the translation quality is good enough to worth post-editing (below 50% TER). The translation quality of the French→English pair was lower than expected. This phenomenon is due mainly to two reasons. First, reviewers employed a two-pass review process generating final translations that significantly differ from those that would be obtained in a single pass as in the rest of translation pairs. Second, we believe that the MT system providing the automatic English translations from French, did not properly adapt to the specific domain of the French courses.

6.3.3 Comparison with mainstream providers

One of the questions that arises is how the adapted systems deployed in this work compare to systems from mainstream providers and, in particular, to state-of-the-art YouTube's automatic captioning and Google Translate systems. To this purpose, a benchmark was defined with videos from MOOCs offered by the EMMA platform. Table 6.3 shows, for each transcribed language, the number of videos included in the benchmark, their duration, and the WER achieved by the TTP and YouTube's automatic captioning. Note that the Italian comparison was presented also in Chapter 5.

From the results in Table 6.3, we can conclude that YouTube's WER is higher than that of TTPs systems for all languages, and more precisely, the relative WER increase over the TTP is nearly 70% on average. The main reason behind these results is the fact that YouTube uses general-purpose ASR systems, while the ASR systems integrated into the TTP are automatically adapted to the task as described in Section 6.2. The English ASR system obtained a surprisingly high WER compared with that reported in Table 6.1 based on the same

Table 6.3: Videos, duration (hrs.), and TTP and YouTube WER for each language.

Language	Videos	Hrs.	TTP	YouTube
Spanish	23	3.5	14.8	22.5
Italian	3	4.0	17.1	31.6
English	9	0.4	39.2	65.9
Dutch	2	1.1	24.5	41.1
French	18	2.3	20.6	32.0

technology. An error analysis on the English videos studied in this work led to the conclusion that the accent of the only speaker in these videos was especially difficult to understand.

The benchmark used to compare transcriptions was enlarged for the purpose of comparing translations. Table 6.4 shows, for each translation pair, the number of videos included in the translation benchmark, their duration, and the TER obtained with the TTP and Google Translate. These videos were previously transcribed in order to be translated. In the case of English into Spanish and French into English, the same English and French videos transcribed in Table 6.3 were then translated.

Table 6.4: Videos, duration, and TTP and Google TER for each translation pair.

Tl. pair	Videos	Hrs.	TTP	Google
Es $\rightarrow$ En	250	13.9	33.9	44.3
En $\rightarrow$ Es	9	0.4	35.8	42.4
It $\rightarrow$ En	11	1.1	33.4	39.2
En $\rightarrow$ It	81	5.4	39.7	43.3
Nl $\rightarrow$ En	2	1.2	42.5	45.0
Fr $\rightarrow$ En	18	2.3	52.8	52.6

A general conclusion that can be drawn from Table 6.4 is that Google Translate's MT systems provide higher translation error than TTPs MT systems, except for French into English in which they obtain similar performance. On average, TER figures achieved by Google Translate are higher than those of TTP by 14% relative. Again, as opposed to the general-purpose MT systems provided by Google Translate, TTP systems are adapted to the domain of the video that is being translated, and thus more accurate results are obtained.

6.4 Reviewing time

The time required for reviewers (e.g. lecturers) to post-edit automatic video transcriptions and translations is measured in terms of RTF [21], which is explained in Section 2.5.

A convenient approach to translate a video is to first produce its source subtitles and then translate them into the desired, target languages. This can be done either manually or automatically, or through a combination of both. In particular, a reasonable pipeline combining both is exposed below. Note that, for text documents, only steps 3 and 4 are applied.

1. Automatic high-quality source subtitles are generated by an ASR system.
2. Reviewed source subtitles are produced by lecturers or other staff.
3. Automatic target subtitles are obtained from reviewed source subtitles by using a MT system.
4. Reviewed target subtitles are finally derived by supervising automatic target subtitles.

In general, manual annotation of speech ranges from 10 RTF, in the case of orthographic transcription [16], to 50 RTF, in which a detailed 4-level speech annotation is performed [6]. Expert transcriptionists can achieve as low RTF as 6 [22], but this is not the profile of our lecturers. In our previous work [21], the RTF for manual (orthographic) transcription attained by lecturers is 10.1 ± 1.8 that matches that reported in [16]. For this reason, we take 10 RTF as a reference review time for transcription.

Regarding the RTF for translation, in contrast to transcription, it is more difficult to establish a single reference RTF, except for the rule of thumb of 2500 words per day, since translation is a more complex task requiring a greater cognitive effort and involving different factors such as source and target languages, degree of expertise and experience of the translator, vocabulary specificity, software tools, etc. Having in mind this limitation, specialist translators achieve fully-manual translating rates ranging from 400 to almost 1000 words per hour [14]. Taking these figures into the UPV media repository in which speakers utter 150 words per minute on average, a specialist translator would be translating at 22.5 RTF in the worst case. In the transLectures project [2], seven hours of videos drawn from the UPV media repository were translated ex novo from Spanish into English by two translators achieving an average RTF of 34.1 ± 11.4 RTF. For the sake of comparison and taking into account the profile of our translators (lecturers), hereafter we consider the RTF of manual translation to be 30 RTF.

6.4.1 Transcription reviewing time

Table 6.5 shows, for each transcribed language, the average WER (copied from Table 6.1) and RTF ($\pm$ std. dev.), and regression models to predict RTF as a function of WER. Three regression models were tried: linear, square root and logarithm. In the case of Spanish, detailed information is provided in Table 6.5 on the adjustment of these three regression models. Also, Figure 6.4 shows a scatter plot of RTF (y axis) versus WER (x axis) for each Spanish video (plotted point) and each adjusted regression model. For the rest of transcribed languages, only the details on the adjustment of the logarithmic model are given in Table 6.5 for brevity.

A first important conclusion from the results on transcription reviewing time is that the availability of automatic transcriptions reduces from one third to two thirds the time devoted

Table 6.5: Average WER and RTF ($\pm$ std. dev.), and regression models to predict RTF as a function of WER, for each language.

Language	WER	RTF	Model	R^2	β	*sig*
Spanish	18.4	3.3 ± 1.2	WER	0.87	0.17	$< 10^{-15}$
			$\sqrt{\text{WER}}$	0.90	0.78	$< 10^{-15}$
			ln WER	0.91	1.17	$< 10^{-15}$
English	21.9	5.3 ± 1.7	ln WER	0.92	1.76	$< 10^{-14}$
Italian	25.7	3.9 ± 1.4	ln WER	0.90	1.20	$< 10^{-6}$
Dutch	29.4	5.8 ± 2.5	ln WER	0.85	1.75	$< 10^{-4}$
French	23.2	6.7 ± 0.8	ln WER	0.98	2.17	$< 10^{-15}$

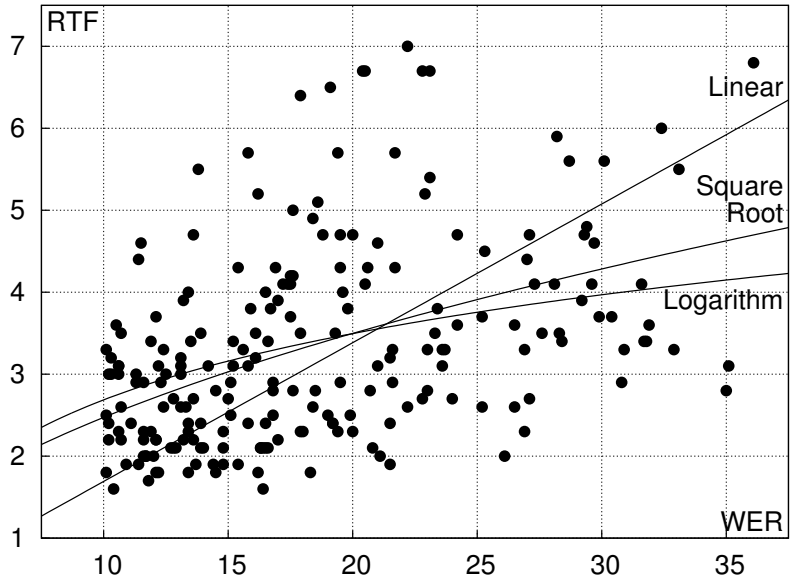

Figure 6.4: RTF as a function of WER for Spanish videos; and three prediction models (linear, square root and logarithm).

to generate video transcriptions. Generally speaking, we may say that the RTF is in between 3 and 7 when starting from automatic transcriptions that are worth post-editing, as discussed in Section 6.3.1. The second important conclusion is that the logarithmic regression model provides a good, statistically significant fit of the observed data, better indeed than the other two models considered. The logarithm model better explains the fact that users tend to ignore automatic transcriptions when the corresponding WER is too high and prefer to retranscribe from scratch to amend a useless automatic transcription. For all languages, the adjustment is

statistically significant ($sig \leq 10^{-4}$) and an important amount of the variability of the data is explained by the model ($R^2 \geq 0.85$).

In a per-language analysis, Dutch presents higher RTF figures than Spanish, Italian and English. We believe this is explained by the interview format of these videos. Finally, the RTF figure for French is not that expected from the WER figure reported, indeed this RTF figure is the highest in the transcription evaluation. The reason behind this RTF figure is the second review process that lecturers carried out in order to guarantee the quality of the video transcriptions for their students. This second review requires at least 1 additional RTF, which is the minimum amount of time needed to check the entire video again.

6.4.2 Translation reviewing time

Table 6.6 shows, for each translation pair, the average TER (copied from Table 6.2) and RTF ($\pm$ std. dev.), and regression models to predict RTF as a function of TER. Translation results are provided in Table 6.6 and Figure 6.5, very much in the same way as above for transcription.

Table 6.6: Average TER and RTF ($\pm$ std. dev.), and regression models to predict RTF as a function of TER, for each translation pair.

Tl. pair	TER	RTF	Model	R^2	β	sig
			TER	0.75	0.25	$< 10^{-15}$
Es→En	33.2	9.1 ± 4.9	$\sqrt{\text{TER}}$	0.80	1.61	$< 10^{-15}$
			$\ln$TER	0.80	2.71	$< 10^{-15}$
En→Es	27.0	7.8 ± 4.9	$\ln$TER	0.82	2.67	$< 10^{-11}$
It→En	37.5	11.3 ± 4.2	$\ln$TER	0.89	3.15	$< 10^{-7}$
En→It	33.8	9.6 ± 5.3	$\ln$TER	0.77	2.76	$< 10^{-15}$
Nl→En	30.7	9.5 ± 3.9	$\ln$TER	0.91	2.89	$< 10^{-2}$
Fr→En	58.9	23.2 ± 8.0	$\ln$TER	0.90	5.67	$< 10^{-4}$

Similarly to transcription, the first important result is that, except for Fr→En, the review time is reduced to approximately one third when the quality of automatic translations is worth post-editing, as explained in Section 6.3.2. The second result is that the logarithmic regression model is among the best explaining the observed data. Again, the logarithmic model better deals with high values of TER to bound the corresponding RTF, since reviewers ignore those automatic translations containing too many errors and prefer to generate the translation from scratch. The amount of variability of the data explained by the model (R^2 values) is not as high as in the review of transcriptions and it is reflected in Figure 6.5 as a greater dispersion of data points. The reason behind this behaviour is the higher complexity of the translation task (compared to transcription) that involves a significant cognitive load.

In a per-translation-pair analysis, the review of Spanish translations from English transcriptions are similar to the translation in the other opposite translation pair, but the RTF figure is even lower for the latter. This fact correlates with the Italian into English and English into

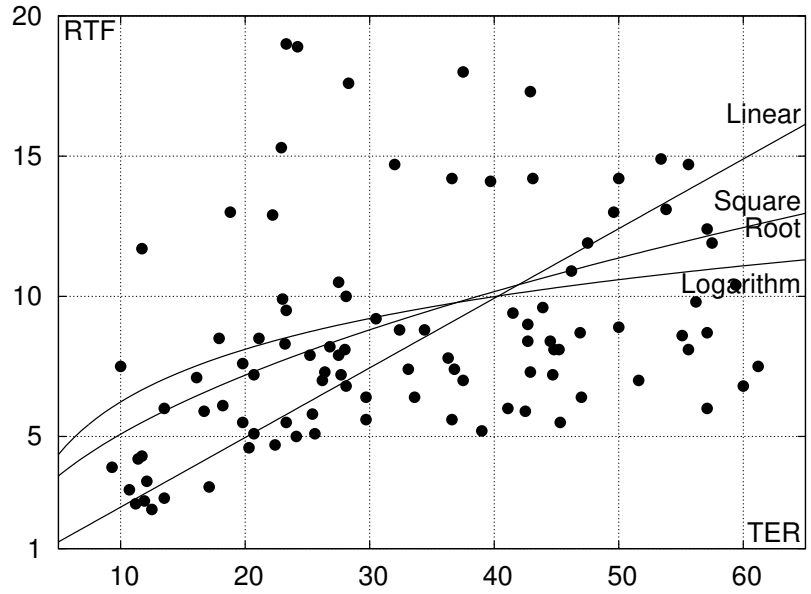

Figure 6.5: RTF as a function of TER for Spanish videos translated into English; and three prediction models (linear, square root and logarithm).

Italian translation pairs, since most of our reviewers are non-native English speakers, and it is easier for them to translate into their mother tongue. The figures for the Dutch into English translation review are very much in line with the previous translation pairs, considering that the quality of the automatic translations was among the best. Finally, the translation of French courses was surprisingly cumbersome, taking far more time than the other translation pairs. This phenomenon is due mainly to two reasons. First, the MT system, that generated the automatic English translations from French, did not properly adapt to the domain of the courses, and second, reviewers employed a two-pass review process that was more costly than the conventional one-pass review process employed in the rest of translation pairs.

6.4.3 Reviewing time across languages

In Section 6.4.1 we have found that, for each language involved, a logarithmic regression model can be adjusted to accurately predict RTF from transcription WER. In Section 6.4.2 we have reached a similar conclusion in translation (i.e. to predict RTF from TER) for each translation pair assessed. Therefore, it is worth asking whether a single logarithmic regression model suffices to accurately predict RTF from WER (TER) across all languages (translation pairs) under study. This is considered in Figure 6.6. The scatter plot at its top shows RTF versus WER, for all languages involved (point symbols), and a single logarithmic regression

model fitted to data (videos) pooled across languages. The scatter plot at its bottom is similar, but for TER.

As for predicting RTF from transcription WER, the fitted logarithmic model shown at the top of Figure 6.6 ($R^2 = 0.87$, $\beta = 1.34$) is statistically significant ($sig < 10^{-15}$). This confirms that reviewing time highly depends on transcription quality and, to a lesser extent, on the language considered. It is worth noting, however, that most data points (videos) are for Spanish (207 out of 277), and thus results are certainly biased towards this language. In this regard, a closer look at the distribution of data points reveals that they are more-or-less clustered by language. This was not unexpected since, after all, there are language and MOOC-dependent factors (e.g. topic, reviewers and reviewing quality requirements) that certainly have some effect on the RTF but fall out of the scope of this work. In any case, the statistical significance of the fit suffices to support the idea that RTF mainly depends on WER, irrespective of the transcription language. For example and to be more precise, taking a couple of reference points on the logarithmic curve we can infer that a one-hour video transcription of 10 WER points will take 3 hours to be reviewed, and a video of the same duration with 20 WER points of transcription error requires almost 4 hours. This is significantly less time than the 10 RTF to transcribe from scratch.

As with WER, the fitted logarithmic model shown at the bottom of Figure 6.6 ($R^2 = 0.78$, $\beta = 2.90$) is statistically significant ($sig < 10^{-15}$) for RTF prediction from TER. As above then, we can confirm that RTF depends more on the translation quality (TER) than on the language pair considered. In contrast to the above results, however, the distribution of data points does not reveal a clear language pair-dependent clustering structure. Taking into account that data points for Spanish (i.e. Spanish→English) are still dominant (250 out of 371), this adds more evidence to support the validity of the fitted logarithmic model. If, for example, a reviewed one-hour video transcription is translated with about 30 TER points, then we may expect an RTF of around 9, that is, 9 hours for reviewing translation. This is obviously much less than the 30 hours (30 RTF) we may expect if translation is carried out manually from scratch; in other words, it entails a reviewing time saving of 70% relative.

6.5 Impact on the case studies

Over the last two years, we have been collecting precise statistics on multilingual data consumption in the two real-life case studies described in Section 1.3: the EMMA platform and the UPV media repository. This data is summarised below in order to better gauge the impact the availability of video transcriptions and translations has had on both case studies.

6.5.1 The EMMA platform

Table 6.7 shows the number of native and non-native students enrolled in the MOOCs offered on the EMMA platform organised by the original language of the course. It goes without

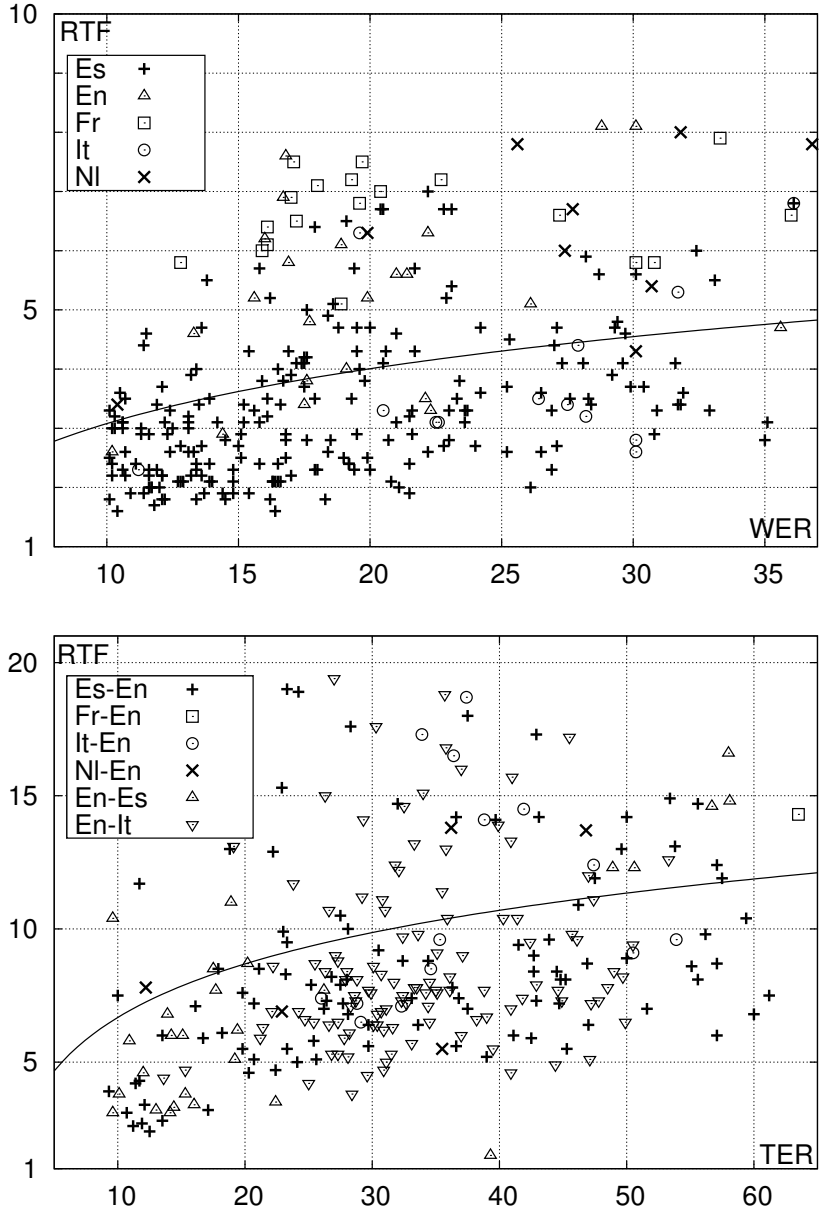

Figure 6.6: RTF versus WER for each language (top) and RTF versus TER for each translation pair (bottom). The curves are logarithmic regression models fitted to data pooled across languages and translation pairs.

saying that non-native students could only follow these MOOCs thanks to the TLP-based multilingual component in EMMA (cf. Section 6.2). The last column in Table 6.7 shows the relative increase in the total number of students over native students due to the enrolment of non-native students.

Table 6.7: Statistics on native and non-native students enrolling in MOOCs on the EMMA platform.

Language	Native students	Non-native students	Rel. Inc.(%)
Spanish	161	547	340
French	983	879	89
Italian	609	259	43
Dutch	501	104	21
English	351	27	8
Total	2605	1816	70

Note that the results in Table 6.7 are given in decreasing order of the relative increment of non-native students. The best results were obtained by MOOCs originally in Spanish and followed by 161 Spanish-speaking students. As these courses were also delivered in English and Italian, 547 non-Spanish-speaking students enrolled in the courses increasing the total number of students by 340% with respect to the Spanish-speaking students. MOOCs in French almost doubled their number of students by offering these courses not only in French, but also in English. MOOCs in Italian and Dutch translated into English also experienced a relative increase in non-native students enrolled of approximately 40% and 20%, respectively. Finally, English courses translated into Spanish had a small relative increase in student enrolment, mainly explained by the fact that English is considered a lingua franca and many non-native students are able to follow the course in English, at least students at this level of education. Overall, the translated versions of the MOOCs facilitated by the TLP in the EMMA platform attracted students that are non-native in the original language of the courses, increasing the total student enrolment by a notable 70%.

Indeed, according to exit questionnaires filled in by almost 1500 students enrolled in EMMA courses, 75% of them appreciated multilingualism as a feature of this platform and 70% found multilingual subtitles useful [7]. Taking into account only those approximately 200 students that replied to mini-questionnaires embedded in 17 running MOOCs, 31% of them always used the translation functionality, that is, the MOOC was originally in a different language from their mother tongue; and 29% of them sometimes used the translation functionality. Indeed, at least 90% of the students always or sometimes using the translation functionality agreed that this functionality enhances the overall value of the EMMA platform and makes EMMA a truly European experience [8].

6.5.2 The UPV media repository

Table 6.8 shows the number of poliMedia videos and subtitle views (in thousands) per language and total from June 2015, when view logs were activated, to May 2016.

Table 6.8: Video and subtitle views (in thousands) per language and total from June 2015 to May 2016.

Video language	Video views	Subtitle views Spanish	Subtitle views English
Spanish	629	6.9	1.1
English	63	1.3	0.5
Total	692	8.2	1.6

The main conclusion that can be drawn from Table 6.8 is that, on average, subtitles were turned on in 1.4% of video views. It is worth noting, however, that a 1.4% of a large number of video views (i.e. almost 700K over the last year) is a significant number of users turning subtitles on (i.e. almost 10K over the last year). Indeed, in relative terms, it is interesting to observe that 2.5% of the English videos had their subtitles activated, in contrast to Spanish videos which did in 1.3% of the views. This results does not come as a surprise since most of our students are native Spanish speakers with English as a foreign language. Finally, Spanish subtitles were predominant when subtitles were activated, being shown in 86% and 71% of the cases for Spanish and English videos, respectively.

Apart from the accessibility benefits for hearing-impaired and foreign students, the availability of transcriptions has allowed the indexing and subsequent search for specific words in such a large video lecture repository. Indeed, this search tool at the UPV media repository allows students to find the specific video clip in which a word is uttered by the lecturer. Thus, students can discard video clips that are not of their interest to focus on those ones in which the concrete concept is explained, saving a significant amount of time. Needless to say that subtitles have also supported students in the arduous note-taking task [15].

6.6 Conclusions

In this chapter, we have reported a large part of the experience we have gained on producing low-cost multilingual video subtitles of publishable quality for MOOCs and OER. Apart from describing the systems, tools and integration components employed for such purpose, a comprehensive evaluation of the results achieved has been provided from three viewpoints: the quality of video transcriptions and translations automatically generated from task-adapted ASR and MT systems, the time required to review them, and the impact multilingual subtitles have had on a MOOC platform and a large video lecture repository.

The quality of automatic transcriptions and translations has been proved to be in most cases below 25% of WER and 50% of TER, respectively. This means that it is worth post-editing them to achieve publishable subtitles instead of generating them ex novo. Indeed, the output of the adapted ASR and MT systems has been positively compared to state-of-the-art automatic transcription and translation tools provided by mainstream providers. More precisely, these systems are on average 38% and 17% better than YouTube's automatic captioning and Google Translate, respectively.

Regarding the review process, we have showed that a lecturer can save from 30% to 70% of the time devoted to review transcriptions, and from 25% to 75% of the translation review time, with respect to performing these tasks from scratch. In addition, a multilingual linear regression model has been proposed to infer the review time (RTF) as a function of WER in the case of transcription, and in terms of TER for translation.

Finally, the availability of multilingual video subtitles has been shown to have a great impact in our case studies. On the one hand, in the EMMA platform, the translation of MOOCs into a second, or even a third language has significantly increased course visibility boosting student enrolment by 70% relative. On the other hand, multilingual subtitles at the UPV media repository have not only improved the accessibility to their video lectures for hearing-impaired and non-native speaking students, but also have allowed the development of added-value functionalities such as indexing and searching capabilities.

Bibliography

[1] The transLectures-UPV Platform (TLP). `www.mllp.upv.es/tlp`, 2016.

[2] Transcription and Translation of Video Lectures (transLectures) project. `translectures.eu`, 2016.

[3] Transcription and Translation Platform (TTP). `ttp.mllp.upv.es`, 2016.

[4] Y. Akita, M. Mimura, and T. Kawahara. Automatic transcription system for meetings of the japanese national congress. In *Proc. of Interspeech*, pages 84–87, 2009.

[5] A. Axelrod, X. He, and J. Gao. Domain adaptation via pseudo in-domain data selection. In *Proc. of EMNLP*, Edinburgh (UK), 2011.

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

[7] C. Ferrari, C. Pennati, A. Tontodonati, K. Tammets, E. Panto, L. Marcellin, and R. Politi. D4.3.2 data and impact analysis report. Deliverable of the European EMMA project, July 2016.

[8] C. Ferrari, C. Pennati, A. Tontodonati, K. Tammets, E. Panto, L. Marcellin, and R. Politi. D4.4 pilot cycle evaluations. Deliverable of the European EMMA project, July 2016.

[9] M. J. F. Gales. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language*, 12(2):75–98, 1998.

[10] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. of Interspeech*, 2007.

[11] T. J. Hazen. Automatic alignment and error correction of human generated transcripts for long speech recordings. In *Proc. of Interspeech 2006*, pages 1606–1609, 2006.

[12] A. Martínez-Villaronga, M. del Agua, J. Andrés-Ferrer, and A. Juan. Language model adaptation for video lectures transcription. In *Proc. of ICASSP*, pages 8450–8454, 2013.

[13] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James. The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, pages 493–502, 2006.

[14] M. Plitt and F. Masselot. A productivity test of statistical machine translation postediting in a typical localisation context. *Prague Bulletin of Mathematical Linguistics*, 93:7–16, 2010.

[15] R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J. P. Robinson, and B. S. Duerstock. Using speech recognition for real-time captioning and lecture transcription in the classroom. *IEEE Transactions on Learning Technologies*, 6(4):299–311, 2013.

[16] D. Reidsma, D. Hofs, , and N. Jovanovic. Designing focused and efficient annotation tools. In L. Noldus, F. Grieco, L. Loijens, and P. Zimmerman, editors, *Proc. of Measuring Behaviour, 5th International Conference on Methods and Techniques in Behavioral Research*, pages 149–152, Wageningen, The Netherlands, 2005.

[17] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas, 2006*, pages 223–231, 2006.

[18] L. Specia. Exploiting objective annotations for measuring translation post-editing effort. In *Proc. of EAMT*, pages 73–80, 2011.

[19] L. Specia, M. Turchi, Z. Wang, J. Shawe-Taylor, and C. Saunders. Improving the confidence of machine translation quality estimates. In *Machine Translation Summit XII*, 2009.

[20] **J**. **D**. **Valor Miró**, P. Baquero-Arnal, J. Civera, C. Turró, and A. Juan. Multilingual videos for moocs and oer. *Educational Technology and Society*, 2017.

[21] **J**. **D**. **Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74:65–75, 2015.

[22] J. D. Williams, I. D. Melamed, T. Alonso, B. Hollister, and J. Wilpon. Crowd-sourcing for difficult transcription of speech. In *Proc. of ASRU*, pages 535–540, 2011.

CHAPTER 7

CONCLUSIONS

Contents

7.1 Summary and future work

In this thesis we studied an efficient way to create multilingual MOOCs and OER by using the ASR and MT technology adapted to the speaker and topic, in order to add subtitles that can be exploited to incorporate added-value functionalities. Searchability, accessibility, translatability, note-taking [1], and discovery of content-related videos, are a few examples of functionalities that are possible due to the existence of subtitles.

To reach this goal, in Chapter 3 we describe the integration of an Spanish state-of-the-art ASR system into the Opencast Matterhorn. The Spanish ASR system was trained using the poliMedia corpus and improved using speaker and topic adaptation. That system obtained good quality transcriptions in our test set, and it is ready to be used in real-life evaluations.

In Chapter 4, our three-phase real-life evaluations discovered that simply by conventional post-editing automatic transcriptions users almost reduced to half the time that would require to generate the transcription from scratch. As expected, this study revealed that the time spent by lecturers reviewing automatic transcriptions directly correlates with the accuracy of said transcriptions. However, it is also shown that the average time required to perform each individual editing operation could be precisely derived and could be applied in the definition of a user model. In addition, the second phase of this study presented a transcription review strategy based on CM and compares it to the conventional post-editing strategy. Finally, a third strategy resulting from the combination of CM with massive adaptation techniques for ASR improved the transcription review efficiency in comparison with the two aforementioned strategies.

In terms of future work, we propose to evaluate alternative variants of intelligent interaction strategies which, while allowing lecturers full control over transcription quality, are better able to exploit confidence measures and visual representation [2, 3]. Also, the most suitable interface design for transcription review could be determined on a case-by-case basis, perhaps as a function of WER. In this scenario, transcriptions with low error rates would be reviewed using an interface that focused user attention on the few words that need correcting, while a conventional post-editing interface would be loaded for transcriptions with higher error rates. However, we also believe that interface design preferences are conditioned by the user profile of our participants. As reported in Chapter 4, lecturers required full control over the final transcription quality, but students or casual users involved in the review process may prioritise the time devoted to review over the transcription quality. This is specially true when dealing with long videos (over 30 minutes) since, as described in the second phase, the possibility of targeting only those segments that have been probably misrecognised becomes more appealing and necessary provided the limited review effort that students or casual users can devote. This latter user profile is better targeted by Serrano [5]. A detailed study of transcription review by students or casual users of longer videos is left as future work.

The next step taken as part of this thesis was to provide multilingual MOOCs and OER at low cost. The final objective was to reach a wider audience by translating MOOCs and OER. In Chapter 5, high-quality ASR and MT systems for a wide range of languages were developed,

evaluated, and subsequently improved, during the entire duration of the EMMA project. The multilingual study of Chapter 6 reported in this thesis reflect that our adapted ASR and MT systems provide draft multilingual subtitles that, in terms of quality, supersede that offered by mainstream providers of this technology as YouTube automatic captioning system and Google Translate. In addition, we show how the access to draft multilingual subtitles allows lecturers to save approximately from 25% to 75% of the time with respect to performing this task ex novo. More precisely, it is shown that draft multilingual subtitles produced by domain-adapted ASR and MT systems reach a level of accuracy that make them worth post-editing, instead of generating them ex novo, saving approximately from 25% to 75% of the time. In addition, the transcription and translation quality of these domain-adapted systems is shown to supersede that offered by mainstream providers as YouTube automatic captioning and Google Translate by about 40% and 20% relative, respectively. Finally, results on user multilingual data consumption are reported from which we can conclude that multilingual subtitles have had a very positive impact in our case studies boosting, in the case of the MOOC platform, student enrolment by 70% relative.

However, we also leave interesting challenges ahead regarding the translation of video lectures and students' interaction. First, the translation of slides integrated into video lectures is an open technical issue still to be tackled. Second, students attending a MOOC in a foreign language need to share their attention between reading the subtitles on the bottom of the screen and looking at the lecturer that is supporting their explanation with the content of the slides. At this point students may find very useful having video lectures automatically dubbed in their mother tongue. Current state-of-the-art text-to-speech technology can provide good quality dubbing at low cost, so that students can devote full attention to what the lecturer is delivering. Finally, the interaction among students via discussion forums in a multilingual MOOC is hampered by the language barrier, creating isolated linguistic communities and preventing students in one language to learn from shared knowledge and experience in other language. Again, the integration of MT technology into discussion forums will lower language barriers from which students will undoubtedly benefit.

Finally, we also have the opportunity to enlarge the scope of this technology by increasing the number of languages and translation pairs that can be automatically transcribed and translated, and extend to related applications like TV programs or films.

7.2 Contributions

It is important to highlight the main contributions of this thesis listing them below.

- To corroborate that massive adaptation techniques in topic and speaker leads to better transcriptions and translations.
- To determine a revision model that combines massive adaptation and user-interaction that is more efficient than the post-editing model.

- To discover that post-editing is the preferred way to review automatic transcription and it is more efficient than performing it from scratch.
- To explain and determine the dependencies between transcription quality and review time devoted by lecturers.
- To develop high-quality efficient ASR and MT systems for multiple languages that have been used to captioning thousands of video lectures.
- To propose a double-selection approach to generate high-quality translation systems without in-domain parallel data.
- To evaluate a real-life scenario for the post-editing of multilingual captions to highlight their benefits.

We also list the scientific publications on which the present thesis is based, in chronological order. First, we present the journal publications associated to this thesis.

- **J. D. Valor Miró**, P. Baquero-Arnal, J. Civera, C. Turró, and A. Juan. Multilingual videos for moocs and oer. *Educational Technology and Society*, 2017. [15]
- **J. D. Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74:65–75, 2015. [17]
- **J. D. Valor Miró**, R. N. Spencer, A. Pérez-González, G. Garcés, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning: The Journal of Open, Distance and e-Learning*, 29(1):72–85, 2014. [20]

Secondly, we detail below the research conference proceedings on which this thesis is based.

- **J. D. Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficient generation of high-quality multilingual subtitles for video lecture repositories. *Design for Teaching and Learning in a Networked World*, pages 485–490, 2015. [18]
- **J. D. Valor Miró**, A. Pérez González de Martos, J. Civera, and A. Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. *Advances in Speech and Language Technologies for Iberian Languages*, pages 237–246, 2012. [16]

Thirdly, we mention some partial contributions to scientific publications that were included in this thesis.

- A. Pérez González de Martos, J. A. Silvestre-Cerdà, **J. D. Valor Miró**, J. Civera, and A. Juan. MLLP transcription and translation platform. In *Tenth European Conference On Technology Enhanced Learning (EC-TEL 2015)*. Universidad Carlos III de Madrid, 2015. [4]

- J. A. Silvestre, M. Del Agua Teba, G. Gascó, A. Giménez, A. Martínez-Villaronga, I. Sanchez-Cortina, N. Serrano Martinez-Santos, **J. D. Valor Miró**, J. Andrés, J. Civera, et al. transLectures. In *IberSPEECH 2012-VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSPEECH 2012)*, pages 345–351. Universidad Autónoma de Madrid, 2012. [6]

Moreover, some educational publications as a result of this work were published.

- **J. D. Valor Miró**, C. Turró, J. Civera, and A. Juan. Generación eficiente de transcripciones y traducciones automáticas en polimedia. In *II Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2016)*. Universitat Politècnica de València, 2015. [22]

- **J. D. Valor Miró**, C. Turró, J. Civera, and A. Juan. Evaluación de la revisión de transcripciones y traducciones automáticas de vídeos polimedia. In *I Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2015)*, pages 463–467. Universitat Politècnica de València, 2015. [21]

- **J. D. Valor Miró**, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para vídeos polimedia. In *I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278. Universitat Politècnica de València, 2014. [19]

Finally, the results and methodologies of this thesis contributed in a significant way to the European projects *EMMA* and *transLectures*, concretely to the work presented in the public reports listed below.

- UPVLC. D3.3.3: Report on evaluation of final transcription and translation models. Technical report, EMMA, 2016. [14]

- UPVLC. D2.3.3: Report on final transcription and translation models. Technical report, EMMA, 2016. [13]

- UPVLC. D3.3.2: Report on evaluation of improved transcription and translation models. Technical report, EMMA, 2015. [12]

- UPVLC. D2.3.2: Report on improved transcription and translation models. Technical report, EMMA, 2015. [11]

- UPVLC. D3.3.1: Report on evaluation of initial transcription and translation models. Technical report, EMMA, 2014. [9]

- UPVLC. D2.3.1: Report on initial transcription and translation models. Technical report, EMMA, 2014. [8]

- UPVLC. D6.3.2: Second report on evaluations at the case studies. Technical report, transLectures, 2014. [10]

- UPVLC. D6.3.1: First report on evaluations at the case studies. Technical report, transLectures, 2013. [7]

Bibliography

[1] J. Glass, T. J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay. Recent Progress in the MIT Spoken Lecture Processing Project. In *Proc. of Interspeech*, 2007.

[2] S. Luz, M. Masoodian, and B. Rogers. Supporting Collaborative Transcription of Recorded Speech with a 3D Game Interface. In *Proc. of the 14th International Conference on Knowledge-based and Intelligent Information and Engineering Systems (KES): Part IV*, pages 394–401, 2010.

[3] S. Luz, M. Masoodian, B. Rogers, and C. Deering. Interface design strategies for computer-assisted speech transcription. In *Proc. of the 20th Australasian Conference on Computer-Human Interaction: Designing for Habitus and Habitat (OZCHI)*, pages 203–210, 2008.

[4] A. Pérez González de Martos, J. A. Silvestre-Cerdà, **J. D. Valor Miró**, J. Civera, and A. Juan. MLLP transcription and translation platform. In *Tenth European Conference On Technology Enhanced Learning (EC-TEL 2015)*. Universidad Carlos III de Madrid, 2015.

[5] N. Serrano, A. Giménez, J. Civera, A. Sanchis, and A. Juan. Interactive handwriting recognition with limited user effort. *International Journal on Document Analysis and Recognition*, pages 1–13, 2013.

[6] J. A. Silvestre, M. Del Agua Teba, G. Gascó, A. Giménez, A. Martínez-Villaronga, I. Sanchez-Cortina, N. Serrano Martinez-Santos, **J. D. Valor Miró**, J. Andrés, J. Civera, et al. transLectures. In *IberSPEECH 2012-VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop (IberSPEECH 2012)*, pages 345–351. Universidad Autónoma de Madrid, 2012.

[7] UPVLC. D6.3.1: First report on evaluations at the case studies. Technical report, transLectures, 2013.

[8] UPVLC. D2.3.1: Report on initial transcription and translation models. Technical report, EMMA, 2014.

[9] UPVLC. D3.3.1: Report on evaluation of initial transcription and translation models. Technical report, EMMA, 2014.

[10] UPVLC. D6.3.2: Second report on evaluations at the case studies. Technical report, transLectures, 2014.

[11] UPVLC. D2.3.2: Report on improved transcription and translation models. Technical report, EMMA, 2015.

[12] UPVLC. D3.3.2: Report on evaluation of improved transcription and translation models. Technical report, EMMA, 2015.

[13] UPVLC. D2.3.3: Report on final transcription and translation models. Technical report, EMMA, 2016.

[14] UPVLC. D3.3.3: Report on evaluation of final transcription and translation models. Technical report, EMMA, 2016.

[15] **J. D. Valor Miró**, P. Baquero-Arnal, J. Civera, C. Turró, and A. Juan. Multilingual videos for moocs and oer. *Educational Technology and Society*, 2017.

[16] **J. D. Valor Miró**, A. Pérez González de Martos, J. Civera, and A. Juan. Integrating a state-of-the-art asr system into the opencast matterhorn platform. *Advances in Speech and Language Technologies for Iberian Languages*, pages 237–246, 2012.

[17] **J. D. Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74:65–75, 2015.

[18] **J. D. Valor Miró**, J. A. Silvestre-Cerdà, J. Civera, C. Turró, and A. Juan. Efficient generation of high-quality multilingual subtitles for video lecture repositories. *Design for Teaching and Learning in a Networked World*, pages 485–490, 2015.

[19] **J. D. Valor Miró**, R. N. Spencer, A. Pérez González de Martos, G. Garcés Díaz-Munío, C. Turró, J. Civera, and A. Juan. Evaluación del proceso de revisión de transcripciones automáticas para

vídeos polimedia. In *I Jornadas de Innovación Educativa y Docencia en Red (IN-RED 2014)*, pages 272–278. Universitat Politècnica de València, 2014.

[20] **J**. **D**. **Valor Miró**, R. N. Spencer, A. Pérez-González, G. Garcés, C. Turró, J. Civera, and A. Juan. Evaluating intelligent interfaces for post-editing automatic transcriptions of online video lectures. *Open Learning: The Journal of Open, Distance and e-Learning*, 29(1):72–85, 2014.

[21] **J**. **D**. **Valor Miró**, C. Turró, J. Civera, and A. Juan. Evaluación de la revisión de transcripciones y traducciones automáticas de vídeos polimedia. In *I Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2015)*, pages 463–467. Universitat Politècnica de València, 2015.

[22] **J**. **D**. **Valor Miró**, C. Turró, J. Civera, and A. Juan. Generación eficiente de transcripciones y traducciones automáticas en polimedia. In *II Congreso Nacional de Innovación Educativa y Docencia en Red (INRED-2016)*. Universitat Politècnica de València, 2015.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AM	Acoustic Model
AI	Artificial Intelligence
AL	Active Learning
ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
CART	Classification and Regression Tree
CAT	Computer-Aided Transcription
CE	Cross-Entropy
CM	Confidence Measures
CMLLR	Constrained Maximum Likelihood Linear Regression
DeX	Docència en Xarxa
DNN	Deep Neural Network
EMMA	European Multiple MOOC Aggregator
GPU	Graphical Processing Unit
HMM/DNN	Hidden Markov Model emiting with Deep Neural Networks
HMM/GMM	Hidden Markov Model emiting with Gaussian Mixture Models
ICT	Information and Communication Technology

INS	Infrequent *n*-gram Selection
LM	Language Model
mDNN	Multilingual Deep Neural Network
ML	Machine Learning
MLLR	Maximum Likelihood Linear Regression
MT	Machine Translation
MOOC	Massive Open Online Course
MMI	Maximum Mutual Information
MFCC	Mel-Frequency Cepstral Coefficient
NLP	Natural Language Processing
OER	Open Educational Resources
OoV	Out-of-Vocabulary
OSGi	Open Services Gateway initiative
OUNL	Open University in the Netherlands
PR	Pattern Recognition
PLE	Personal Learning Environment
RTF	Real Time Factor
RNNLM	Recurrent Neural Network Language Model
RPROP	Resilient Backpropagation
SOA	Service-Oriented Architecture
SGD	Stochastic Gradient Descent
TEL	Technology Enhanced Learning
TER	Translation Error Rate
TLP	transLectures-UPV Platform
TLK	transLectures-UPV Toolkit
TTP	Transcription and Translation Platform

UAB	Universidade Aberta
UPV	Universitat Politècnica de València
UNINA	Università degli Studi di Napoli Federico II
WER	Word Error Rate