Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2348–2357
Marseille, 11–16 May 2020

Subjective evaluation of comprehensibility in movie interactions

Estelle Randria[1,2], Lionel Fontan[2], Maxime Le Coz[2], Isabelle Ferrané[1], Julien Pinquier[1]
[1]IRIT, CNRS, Université Paul Sabatier, Toulouse, France
[2]Archean Labs, 20 place Prax-Paris, 82000 Montauban, France
{estelle.randria, isabelle.ferrane, julien.pinquier}@irit.fr, {lfontan, mlecoz}@archean.tech

Abstract

Various research works have dealt with the comprehensibility of textual, audio, or audiovisual documents, and showed that factors related to text (e.g., linguistic complexity), sound (e.g., speech intelligibility), image (e.g., presence of visual context), or even to cognition and emotion can play a major role in the ability of humans to understand the semantic and pragmatic contents of a given document. However, to date, no reference human data is available that could help investigating the role of the linguistic and extralinguistic information present at these different levels (i.e., linguistic, audio/phonetic, and visual) in multimodal documents (e.g., movies). The present work aimed at building a corpus of human annotations that would help to study further how much and in which way the human perception of comprehensibility (i.e., of the difficulty of comprehension, referred to in this paper as *overall difficulty*) of audiovisual documents is affected (1) by lexical complexity, grammatical complexity, and speech intelligibility, and (2) by the modality/ies (text, audio, video) available to the human recipient. To this end, a corpus of 55 short movie clips was created. Fifteen experts (language teachers) assessed the overall difficulty, the lexical difficulty, the grammatical difficulty and the speech intelligibility of the clips under different conditions in which one or more modality/ies was/were available. A study of the distribution of the experts' ratings showed that the perceived difficulty of the 55 clips range from very easy to very difficult, in all the aspects studied except for the grammatical complexity, for which most of the clips were considered as easy or moderately difficult. The study reflected the relationship existing between lexical complexity and difficulty, grammatical complexity and difficulty and speech intelligibility and difficulty, as lexical complexity and speech intelligibility are strongly and positively correlated to difficulty and the grammatical difficulty is moderately and positively correlated to difficulty. A multiple linear regression with difficulty as the dependent variable and lexical complexity, grammatical complexity and intelligibility as the independent variable achieved an adjusted R^2 of 0.82, indicating that these three variables explain most of the variance associated with the overall perceived difficulty. The results also suggest that documents were considered as most difficult when only the audio modality was available, and that adding text and/or video modalities allowed to decrease the difficulty, the difficulty scores being minimized by the combination of text, audio and video modalities.

Keywords: corpus, movie, comprehensibility, multimodality, linguistic complexity, grammatical complexity, speech intelligibility

1. Introduction

The development of Internet and media service platforms has a major impact on the spread of audiovisual contents. Content understanding is an important issue for the accessibility of massive collections of movie clips. Audiovisual collections can be classified by genre, for example, and it is possible to find labels defining the minimum age required to watch some videos, but a classification to define for which public the content is understandable don't exist. As audiovisual contents are becoming more and more accessible, it is interesting to study if this type of classification is possible. In fact, a document may be perceive as more or less understandable depending on the person age, proficiency or native language... This paper focuses on audiovisual content understanding, more specifically on the difficulty of documents in term of comprehensibility (later referred as overall difficulty). Our goal is to study how humans perceive the overall difficulty of audiovisual contents as well as three other aspects which may contribute to its evaluation: the grammatical and vocabulary complexity and the difficulty to understand speech, which will be referred later as speech intelligibility. Based on human evaluations, this study aims to determine what are the main sources of overall difficulty between the lexical complexity, the grammatical complexity and the speech intelligibility, and between text, audio and video modalities.

2. Previous researches

This section will focus on the previous studies made concerning the understanding of audio content, text content and video content.

2.1. Text content understanding

Trying to evaluate the overall difficulty of a text is a subject which has interested teachers and researchers since long ago (Sherman, 1893; Thorndike, 1921). These studies are mainly led to ensure that people are exposed to texts which are appropriate according to their skills in a given language. This is important specially for teaching purposes: assessing how much a text is difficult to understand is crucial to determine the linguistic proficiency level required to use a text in class. Having an information about the proficiency level needed is important either for native and foreign languages. These researches are useful for native speakers, as children do not have the same mastering of the language than teenagers or adults, but also for non-native speakers, as a beginner do not have the same linguistic skills as more experimented speaker. One application example is the collecting of texts of the same overall difficulty level for handbooks design, another possible application is the redaction of clear and efficient instructions for technical manuals: it is important that instructions are accessible for a large panel of persons.

A text can be read but also heard, so two dimensions of text

understanding will be discussed: reading comprehension and listening comprehension.

2.1.1. Reading comprehension

Reading materials difficulty can be studied from three aspects: readability, legibility and comprehensibility. Readability refers to the study of the linguistic and conceptual aspects of a text to determine its overall difficulty, legibility focuses on the effects of text design (layout, typography...) on comprehension, and comprehensibility refers to the evaluation of text overall difficulty according to a reader's profile. As this study will partly focus on linguistic aspects influence on audiovisual content understanding, we will only deal with text readability. The readability of text can be either measured or predicted. Measuring the readability consists in asking to experts (e.g., teachers) or to a representative group of readers to judge the overall difficulty of given texts. One potential issue with this approach is that the results may not be reproducible as subjects are not systematically asked to justify their choices.

Predicting text readability consists in building formula in "order to provide an index of probable overall difficulty for readers " (Klare, 1974). Sherman (1893) was one of the first researcher to study literature from a statistical point of view, his researches highlighted the existing relationship between length sentences and readability. Later, studies were conducted on the influence of vocabulary on readability. Thorndike (1921) created the first list of words in English, sorted by frequency (around 10,000 words). His studies were motivated by observations first made by German and Russian language teachers: the more frequent a word is used, the easier it is to use.Lively and Pressey (1923) built the first children's readability formula, using Thorndike list. Then, other formulae using lexical and syntactic features were developed. Noticeable ones are Flesch (1948) formula or Dale and Chall (1948) formula. Concerning French, the first formula specific to French language was built by Henry (1975). In his formula he used the number of words per sentence, the ratio between the different words and the total number of words in text, the percentage of words which are not in the Gougenheim (1964) list, the percentage of dialog personal pronouns and the number of *dialog indicators*, a dialog indicator being either an exclamation mark, quotation marks opening dialogues or the first names used alone.

The development of automatic language processing techniques and machine learning techniques contributed to the renewal of the interest in readability. The new machine learning techniques allowed the use of more variables and to substitute previous solutions: for example, Collins-Thompson and Callan (2005) showed that it was possible to replace the list of the usual words by language models, which are based on a probabilistic description of language phenomena. For French, François (2009) tested multiple linear regression, bagging and boosting to perform an automatic estimation of the text difficulties. He used 20 predictors: the mean number of letters per words, the mean length of sentences, Henry dialog indicators, the lexical frequency (measured with a language model) and 11 variables linked to the verbal complexity. François and Fairon (2012) later pursued the previous research, testing more statistical models (including Support Vector Machine and classifier trees) and using new predictors, adding syntactic features and also features linked to semantic information.

These studies showed that the linguistic proficiency, specially the lexical coverage, of the reader were important for reading comprehension. They also highlighted the important role played by lexical and grammatical features (sentence length, verbal complexity...) to predict readability.

2.1.2. Listening comprehension

The factors affecting listening comprehension have been studied for native (L1) and non-native (L2) speakers. It is reasonable to suppose that, like for written texts understanding, the ability of decoding speech sentences will depend on the vocabulary and grammatical knowledge acquired by the listener. For both L1 or L2 listening comprehension, linguistic knowledge does play a major role (Buck, 2001; Anderson, 2005). Vocabulary and grammatical knowledge are needed in order to decode words and sentences and to infer them meaning. Nation (2006) established a relationship between vocabulary knowledge and listening comprehension and Nissan et al. (1995) found that word frequency was linked to item overall difficulty, as the increase of infrequent words rises the difficulty to understand an item. Vocabulary is an important predictor for listening comprehension, this has been identified as a factor of difficulty by L2 students themselves (Goh, 1994). Grammatical knowledge also has a relation with listening comprehension (Carrow-Woolfolk, 1999; Tunmer, 1989).

Cognitive aspects play an important part in L1 and L2 listening: Vandergrift (2007) insisted on the importance of compensatory strategies (like the use of contextual information, visual information or common sense) for an efficient understanding of spoken messages. Because compensatory strategies used for L1 comprehension are useful for L2 listening, it is important to develop them in the native language as the same strategies can be used for foreign languages. The working memory, which is the ability of storing and manipulating information, is also crucial, either for text comprehension (Daneman and Merikle, 1996), vocabulary (Stokes and Klee, 2009), grammatical knowledge (Robinson et al., 2003) and listening comprehension (Florit et al., 2013). The working memory allows to store, concatenate and articulate the input information simultaneously through all the listening process. Finally, inference, theory of mind and comprehension monitoring are primordial to allow a listener to interpret messages coherently ((Perfetti and Stafura, 2014; Kim and Phillips, 2014): a text alone is sometimes not enough to fully understand the meaning as some information may be missing if the listener lacks some background knowledge. The listener has to identify incoherence (comprehension monitoring), he has to give meaning to what is not explicitly said in the text (inference) and he has to understand other mental states and behaviours (theory of mind). Theory of mind is important because a certain level of reasoning can serve to understand the speakers behaviour and intentions and then gives more cues on what is happening. The combination of all these cognitive abilities are needed to succeed in listen-

ing comprehension, from decoding to interpreting what is being heard.

Another aspect which can affect listening comprehension is the affective dimension. In educational environments, the listening comprehension task can tend to be stressful for students because they are confronted to the fear of failing. Thus, they might succeed poorly in the listening comprehension task. Elkhafaifi (2005) and Aneiro (1989) highlighted the influence of students anxiety and of a stressful environment on listening comprehension success; Noro (2006) explained that listening overall difficulty can lead to negative emotional reactions such as a lack of attention or a low self-confidence. If these negative emotions are not handled correctly by the students and the teachers, listening tasks can tend to be less successful. Reducing students anxiety, for example by developing affective methods (*e.g* help between students, tools to ease the comprehension), can improve their ability to understand speech (Kurita, 2012).

In short, succeeding in listening comprehension depends both on the linguistic features of speech and on the linguistic knowledge of the listener, the cognitive mechanisms developed by the listener are very important to succeed in listening and his anxiety towards the listening comprehension task can influence his performances.

2.2. Audio content understanding

The studies on audio content understanding are made in various fields. Some researches are also led for teaching purposes (they often focus on what makes listening difficult for language learners) while others focus on the audio signal quality or on the speech intelligibility of speakers. The aim of this studies can be to ensure emergency messages clearness in public places or to improve audio signal quality where exchanging messages through audio devices is mandatory, like for the pilots communicating with the persons based on the control towers.

In the previous part, some points that made listening difficult for native and non-native speakers were evoked: the linguistic features, the cognitive and the affective aspects. Here the focus will be made on auditory features of speech. The way of delivering a message is important for listening comprehension. For English, prosody tends to be an important factor: stress and intonation can be important in order to understand a message (Wong and Waring, 2010), for instance, stress can be used to emphasize important words. Aside from prosody, the speed delivery or the accent of the speakers can affect the comprehension. Teachers in Boyle (1984) study considered that the way a message is produced (clarity, quality, accent...) is important for listening comprehension. Goh (1994) and Hayati (2010) observed that a high speech rate could be a source of overall difficulty for the students, even more if they are not used to listening to the language; speech rate preferences can be related to other factors like the listener proficiency, the speaker pronunciation, or the listener personal references (Zhao, 1997). Pauses and hesitations (also named disfluencies) can have different effects for native and non-native speakers. For L1 listeners, hesitations and pauses can be used as supplementary information, which can help to emphasize some words or the speaker intentions (Corley and Hartsuiker, 2003).Blau (1991) found that pauses could provide either help or confusion to non-native listeners, while Voss (1979) concluded that pauses and hesitations brought perceptual barriers to second-language learners. No conclusions can be made on the pauses and hesitations influence on non-native listening comprehension: pauses and hesitations can be either beneficial or distracting.

Chang and Read (2008) studied the impact of unfamiliar accents on listening comprehension, their studies concluded that the comprehension is only affected if the listener is not used to hearing the accent. This problem affects evenly native speakers Ikeno and Hansen (2006) as Weil (2003) demonstrated that native speakers perform better on understanding once they were familiar with a person accent.

The sound environment can also be perceived as a factor of overall difficulty. This is mentioned by teachers in Boyle (1984) study, this study was aimed to list the factors affecting L2 listening comprehension from teachers and students point of view. Adank et al. (2009) and Larsby et al. (2005) explained that native listeners also encountered difficulties to understand what is being said if the circumstances were not ideal. The effects of adverse conditions on listening remain harder for non-native listeners, because they have to face difficult hearing conditions and their own imperfect knowledge of the listened language: they tend to be more sensitive to increasing noises, babbles behind initial speech or reverberations. But bilingual listeners can be less affected by noisy conditions (Lecumberri et al., 2010). In this part, it was seen that prosody, speech rate, pauses and hesitations, accents and a deteriorated sound environment have an influence on audio content understanding.

2.3. Video content understanding

Concerning video content understanding, studies may be scarcer, but some can be found concerning the cognitive load: does the quantity of information influence the understanding? Other studies exist in language learning field, to determine the sources of difficulty or the facilitators of video understanding, or what is more efficient for language learning, between audio and video movie clips.

2.3.1. Subtitles effects on understanding

The influence of subtitles for video content understanding is a subject of interest for language learning purposes. In order to define whether it is beneficial or detrimental, several studies have been made to compare results of L2 students exposed to subtitles and L2 students who don't have subtitles. Two types of subtitles for language learning exist: L1 subtitles, where the subtitles are in the native language of students and L2 subtitles where subtitles are on the language being learnt. Studies mainly prove that subtitles (L1 or L2) have a positive effect on understanding for second language learners: they perform better on comprehension tests (Perez et al., 2013). L1 subtitles are appropriate for beginners as they have less language proficiency, but L2 subtitles bring more effective results as there is a redundancy between what is being said and what is being read (Hayati and Mohmedi, 2011). Mitterer and J.McQueen (2009) study confirmed that native language subtitles harmed the understanding process of non-native speakers, while foreign sub-

titles were helpful. Other studies took interest in the effect of subtitles in other scopes than language learning, and it was found that, in the same way, subtitles brought positive effects for understanding a video content, either in native or foreign language (Markham et al., 2001).

In general, studies show that subtitles are beneficial for understanding, but some studies imply that subtitles may bring too much cognitive load to the viewer. In fact, it can be feared that subtitles may increase cognitive load leading to a decrease in the understanding. Kalyuga et al. (1999) led a study which showed that subtitles could bring cognitive overload and lower performances. But recent research on the impact of subtitles on cognitive load found that subtitles did not bring cognitive overload (Kruger and Matthew, 2013). It is difficult for us to say what are the real effects of subtitles on cognitive loads.

2.3.2. Visual cues influence on content understanding

It has been demonstrated that gestures are crucial for human communication and are beneficial for speaking and learning languages (Goldin-Meadow and Alibali, 2013; Kellerman, 1992). Dahl and Ludvigsen (2014) led a study proving that gestures facilitated native and foreign language listening comprehension: even if non-native and native speakers won't use the gestures information the same way, it will help them to reach a better understanding. It can be supposed, from these observations, that gestures could bring a better understanding of video content. In Sueyoshi and Hardison (2005), it was shown that the access to video with gestures and visual cues helped second language learners to perform better on comprehension tasks. For learners with higher proficiency they performed better with facial cues, the lower-level learners performed better when they had access to the facial cues and the gestures simultaneously. Harmer (2007) emitted the opinion that video materials bring benefits for foreign language learning, thanks to the access to facial expressions, gestures and other visual cues.

In this part, it was explained that subtitles and visual cues have an influence on understanding. But contrary to audio and text content, no studies seem to have been led about the influence of elements like linguistic features or speech production on video content understanding. What have been studied, on the other hand are the effects of video material on language comprehension, and they were compared to audio and text materials. All the researches on this matter have shown that video is more efficient for language comprehension (Jones and Plass, 2002; Batel, 2014; Yasin et al., 2018). These studies led us to make the assumption that videos are easier to understand than audio movie clips, which could explain the differences in their effects on language learning success.

Researches on content understanding have shown that linguistic features have an influence for both reading and listening, the way of producing speech (prosody, speech rate) and the sound environment play a role for audio content comprehension. For video content, the presence of subtitles and visual cues appear to be facilitators for understanding. But apart from these aspects, which are inherent to the movie clips themselves, the profile of the person reading, listening or viewing, has a major role in the comprehension: the proficiency in the targeted language, the cognitive skills developed and how comfortable the learner feels with the listening task are as important for a successful understanding. The study focuses on the inherent features of movie clips that affect understandings. It was decided to study the lexical complexity, the grammatical complexity and the speech intelligibility (speech intelligibility is how comprehensible the speech is, it encompasses the speech production quality and the environment effects on speech comprehension) and their influence on the perception of overall difficulty.

3. Material and methods

3.1. Corpus

The study consists in analyzing the human perception of audiovisual movie clips overall difficulty. An appropriate corpus to lead the research was needed, so that they could evaluate the overall difficulty depending on the available modalities. In this research three modalities are exploited: the audio modality, the video modality and the text modality. A corpus composed of 55 clips of 15 popular French movies was built, the movies were selected to provide various genres, release dates, sound qualities and different French language varieties and linguistic levels. The focus was made on interactions as defined by Traverso (2013). Understanding contextual elements (who is speaking to whom, when, how, why...) is important to understand the communication situation. Considering interactions as understanding units is a first step towards the evaluation of difficulty. Thus, full-action scenes and scenes without speech were excluded as there were considered as not relevant enough to assess overall difficulty. The clips were chosen to contain long enough interactions between the characters. Three to five clips per movie were selected on which the video, audio and transcription content were extracted. The corpus is composed of 55 movie clips (2541 seconds) each one being available under three different formats:

- text: movie clips exact transcriptions (7225 words),
- audio: clip sound track only,
- video: which includes both audio and images.

Table 3.1. lists all the movies used for the corpus creation, their year of release and the number of movie clips included in the final corpus.

3.2. Participants

We chose foreign-language teacher as experts for our study, because foreign-language teachers often face the problem of evaluating if a document (whether in text, audio, or audiovisual form) is appropriate for their students in terms of overall difficulty. Fifteen teachers of French as a foreign language were recruited based on them matching the following criteria:

- Native French speakers;
- With at least three years of teaching experience with learners of various proficiency levels in French;

Title	Year	Extracts
Le fabuleux destin d'Amélie Poulain	2001	3
Cyrano de Bergerac	1990	4
Delicatessen	1991	3
Embrassez qui vous voudrez	2002	4
Intouchables	2011	3
La chèvre	1981	4
La folie des grandeurs	1971	3
La gloire de mon père	1990	5
La grande vadrouille	1966	4
Le petit Nicolas	2009	4
Les choristes	2004	4
Les plages d'Agnès	2008	3
Qu'est-ce qu'on a fait au bon Dieu	2014	3
Séraphine	2008	4
Un long dimanche de fiançailles	2004	4

Table 1: Movies of the corpus

- Familiar with the use of audiovisual movie clips in class;
- Without any (self-reported) hearing impairment.

The 15 teachers –13 females; age range: 27-63 years; mean age: 37 years –had a teaching experience ranging from 3 to 40 years (mean: 11 years; standard deviation: 9 years). All teachers received monetary compensation for their participation.

3.3. Rating procedure

An online graphical user interface (GUI) was developed to present the 55 movie clips to each participant using five different (combination of) modalities:

- text: only the exact transcript of the clip was available to the participant;
- audio;
- audio + text;
- audio + video;
- audio + video + text.

For each participant, each clip was presented in a single of the modalities listed above, for a total of 11 text presentations, 11 audio presentations, 11 audio + text presentations, 11 audio + video presentations, and 11 audio + video + text presentations. At the end of the rating procedure, the 55 movie clips were presented in each of the five presentation modality to exactly three participants.
The participants completed the test online with their own hardware. They were instructed to conduct the experiment in a quiet room, using a PC and headphones.
The movie clips were presented to each participant in a random order. For each movie clip, the GUI presented a text area and an audio/video player, as well as three-to-four sliders depending on the presentation modality. Participants were instructed to use the sliders to rate each clip in terms of overall difficulty (from 0 –very easy to 100 –very difficult), lexical complexity (from 0 –very easy to 100 –very difficult), grammatical complexity (from 0 –very easy to 100 –very difficult), and, for each presentation modality in which the audio was available, speech intelligibility (from 0 –totally intelligible to 100 –totally unintelligible). The position of each slider was initialized with a central value of 50. The GUI also provided input text areas in order to allow the participants to leave comments in order to justify the adjustment of each slider. These comments were mandatory for ratings of overall difficulty, and optional for ratings of lexical complexity, grammatical complexity, and speech intelligibility.
Prior to the proper rating task, each participant was familiarized with the procedure by rating a set of five training movie clips. Participants were not required to rate all 55 movie clips in a single rating session: they could stop whenever they required to take a break.

4. Results

In this section, descriptive statistics will be provided and the focus will be made on the distributions, the variation intervals of the ratings and on the movies ranking depending on the mean overall difficulty. The next part will deal with the influence of lexical complexity, grammatical complexity and speech intelligibility on the overall difficulty. The last part will focus on the influence of modalities on overall difficulty and speech intelligibility.

4.1. Descriptive statistics

4.1.1. Distributions of the ratings

The histograms of the distribution of the overall difficulty, the lexical complexity, the grammatical complexity and speech intelligibility are shown in figures 1 to 4.

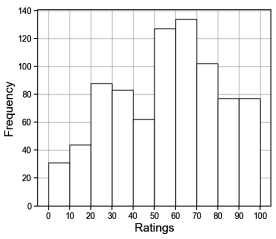

Figure 1: Distribution of the ratings of overall difficulty

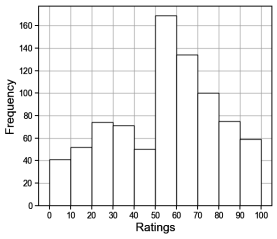

Figure 2: Distribution of the ratings of lexical complexity

For all rating dimensions, the participants used the full range of the evaluation scale, suggesting that the corpus was rather diverse in term of overall difficulty, lexical complexity, grammatical complexity and intelligibility. Overall difficulty and lexical difficulty have a median of respectively 58 and 55 indicating that the majority of the clips were considered rather difficult lexically speaking and overall. On the other hand, the median of the grammatical complexity is 36 and therefore indicates a usage of fairly simple grammar through the corpus. With a median of 50, the intelligibility score is rather balanced.

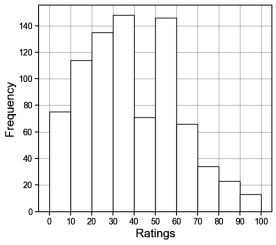

Figure 3: Distribution of the ratings of grammatical complexity

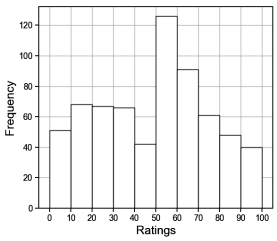

Figure 4: Distribution of the ratings of speech intelligibility

4.1.2. Ranking of the movies as a function of their average overall difficulty

To describe further the corpus, an analysis of how the average overall difficulty varies depending on the movies was made. The bar plots in Figure 5 and Table 2 give the ranking of movies according to the average overall difficulty, sorted from the easiest (1^{st}) to the most difficult (15^{th}).

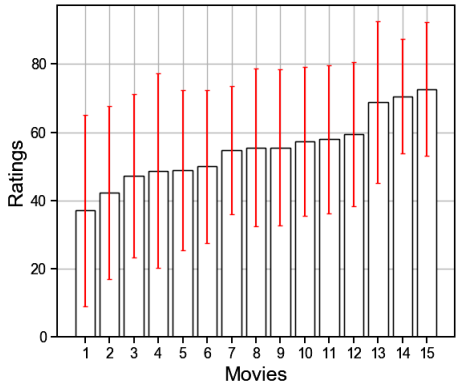

Figure 5: Average overall difficulty as a function of movies. Error bars represent $\pm$ 1 standard deviation.

The bar plots in Figure 5 show that the movie corpus includes easy to difficult movies as the average overall difficulty ranges between 37 and 72. A closer look at the values show that the average overall difficulty is inferior to 50 for six movies, and superior to 50 for nine movies, which means that the corpus is rather balanced in terms of overall difficulty.

4.2. Relationship between overall difficulty and lexical, grammatical and speech intelligibility ratings

In this section, the relationship between ratings of overall difficulty and sub-ratings (lexical and grammatical complexity, speech intelligibility) is investigated through bivariate correlations and multiple linear regressions.

4.2.1. Bivariate correlations

Table 3 presents the results of Spearman correlations computed between ratings of overall difficulty and the three sub-ratings. In the case of speech intelligibility, only the ratings given when audio information was available to the raters were taken into account.

Ranking	Movie
1	Le petit Nicolas
2	Le fabuleux destin d'Amélie Poulain
3	Les choristes
4	La gloire de mon père
5	Les plages d'Agnès
6	Séraphine
7	Intouchables
8	La folie des grandeurs
9	La chèvre
10	Embrassez qui vous voudrez
11	La grande vadrouille
12	Qu'est-ce qu'on a fait au bon Dieu
13	Cyrano de Bergerac
14	Un long dimanche de fiançailles
15	Delicatessen

Table 2: Ranking of the movies as a function of their average overall difficulty

	Lexic. comp.	Gram. comp.	Speech intel.
Overall difficulty	0.74***	0.56***	0.63***

Table 3: Spearman correlations between overall difficulty and lexical complexity (Lexic. comp.), grammatical complexity (Gram. comp.), and speech intelligibility (Speech Intel.) (*** $p \leq 0.001$)

Highly significant, moderate-to-strong positive correlations are found between the overall difficulty and all of the three sub-ratings. This indicates that, as expected, (1) the more complex the lexicon and the grammar are perceived, the more difficult the movie clip is rated, and (2) the higher the speech intelligibility, the easier it is rated. In order to investigate further the relationship between the overall difficulty and the three sub-ratings, multiple linear regressions were computed.

4.2.2. Multiple linear regressions

Overall difficulty and sub-ratings were averaged for each movie-clip and presentation modality. Two multiple linear regressions (MLR) were then performed. A first MLR was computed with the overall difficulty as the dependent variable and the lexical and grammatical complexities as independent variables.

This MLR achieved a high correlation with the predicted overall difficulty, with an adjusted R^2 of 0.76. The non-standardized coefficients (NsCoef) show that the lexical complexity (NsCoef = 0.69) holds more weight in the regression than grammatical complexity (NsCoef = 0.34). The Figure 6 presents a scatterplot relating the predicted overall difficulty to the average human ratings of overall difficulty.

Then, to check if taking into account speech intelligibility can improve the prediction, another MLR was computed,

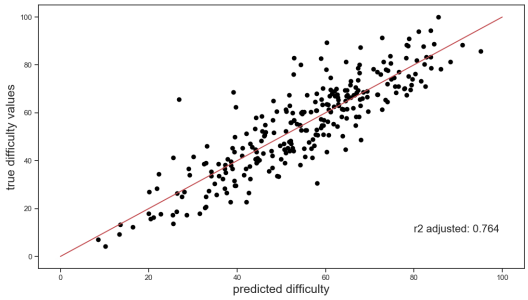

Figure 6: Scatterplot relating human ratings of overall difficulty to predicted ratings of overall difficulty, using a multiple linear regression with lexical complexity and grammatical complexity as independent variables

with the overall difficulty as the dependent variable, and lexical complexity, grammatical complexity, and speech intelligibility as independent variables. For this computation, ratings obtained when only textual information were available to the raters were discarded.

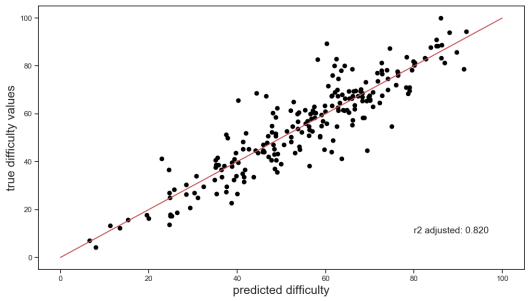

Figure 7: Scatterplot relating human ratings of overall difficulty to predicted ratings of overall difficulty, using a multiple linear regression with lexical complexity, grammatical complexity, and speech intelligibility as independent variables

Adding speech intelligibility to the independent variables in a second MLR allowed to obtain a better prediction than with lexical and grammatical complexity only, with an adjusted R^2 reaching a value of 0.82. The lexical complexity has still the strongest weight for the prediction of overall difficulty (NsCoef = 0.55), followed by grammatical complexity (NsCoef = 0.31), and speech intelligibility (NsCoef = 0.28). The relationship between predicted and actual overall complexity can be visualized in Figure 7.

4.3. Influence of modalities

Another interesting aspect of this study was to observe the rating variation depending on the available modalities. It can be expected that:

- the overall perceived difficulty will be higher for the audio only (A) condition than for the audio+text (AT) condition and for the audio+video (AV) condition;
- the overall perceived difficulty will be higher for AV and AT conditions than for audio+video+text (AVT) condition;
- speech intelligibility will increase if the video and/or text modality are combined with audio modality.

No relationships are expected between modalities and lexical or grammatical complexity.
For each modalities presentation and for overall complexity and each sub-ratings, the ratings mean and standard deviation were calculated, to visualize the evolution of ratings depending on available modalities. The results for overall difficulty and speech intelligibility are in Figures 8 and 9.

4.3.1. Influence of modalities on overall difficulty

First, it can be noticed that the mean of overall difficulty ratings is the highest for audio modality alone: in the corpus, the movie clips with audio only were the most difficult to understand. As foreseen, the means are lower for AV and AT conditions than when audio modality is alone. The lowest mean of overall difficulty is obtained with the AVT conditions, as supposed, it occurs that for this study corpus, having the three modalities help to minimize the perceived overall difficulty.

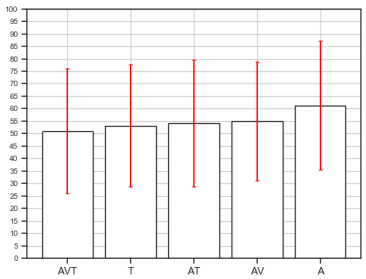

Figure 8: Mean ratings of overall difficulty as a function of the modality of presentation (A: audio alone; AV: audio + video; AT: audio + text; AVT: audio + video + text). Error bars represent $\pm$ 1 standard deviation.

4.3.2. Influence of modalities on speech intelligibility

To both watch the video and the text should be a plus, if the video is not enough to disambiguate what is being said, the text definitely eliminates this problem (even if it can feared that too many information may lead to a cognitive overload). This theory can be supported by the observations made on the evolution of the mean of speech intelligibility ratings depending on the modalities. Figure 9 shows that movie clips with audio only were perceived as less intelligible by the participants and that adding video and/or text modality improved speech intelligibility. The most intelligible movie clips were the ones where both audio, video and text were available. According to the bar plots, the less intelligible movie clips are when audio is alone, the more intelligible movie clips are when there is the audio+video+text condition, complementing audio with other modalities improves speech intelligibility.

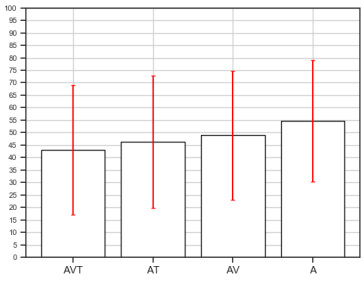

Figure 9: Mean ratings of speech intelligibility (lower ratings = higher intelligibility) as a function of the modality of presentation (A: audio alone; AV: audio + video; AT: audio + text; AVT: audio + video + text). Error bars represent $\pm$ 1 standard deviation.

5. Discussion and conclusions

Focusing on audiovisual content understanding, we built a corpus based on subjective evaluations of 55 clips extracted from 15 movies. The evaluations were realized by 15 teachers who were used to the task of comprehension difficulty evaluation, we analyzed their ratings considering different points of view: difficulty levels and modalities. The movie clips corpus has proven to be rather well balanced, as it contains movie clips with variable lexical complexity, overall difficulty and speech intelligibility. But the distribution of the ratings highlighted a tendency from the participants to assess the grammatical complexity as easy. Some studies showed that oral passages with a high orality (the extent to which passages are similar to spoken language) often have a simpler syntax (Tannen, 1982). As the corpus is composed of movie clips, the orality is more likely to be high. This could explain why the grammatical complexity ratings tend to be lower than the other ratings.

Concerning the influence of lexical and grammatical complexity and speech intelligibility, at the corpus level, it was confirmed that a positive correlation did exist between these factors and the overall difficulty. Performing a multiple linear regression, which considered overall difficulty as a dependent variable and both lexical and grammatical complexity and speech intelligibility as independent variables, confirmed that the combination of this three variables allowed a satisfying prediction of overall difficulty, with an adjusted R^2 reaching 0.82. This study took in consideration the most important factors explaining the variance associated with the overall perceived difficulty for audiovisual content. By using efficient automatic estimators for lexical complexity, grammatical complexity and speech intelligibility, it should be possible to predict overall difficulty. The adjusted R^2 not reaching a value of 1 means that some elements may be missing to perform a perfect prediction. It is possible that minor factors impacting overall difficulty (linked to cognition, for example) were not included in the study even if they are not primordial for the prediction. Also, as the evaluations were performed by humans, some additional factors, inherent to the participants may have influence the ratings, bringing a supplementary margin of error: for example, it is possible that they did the evaluation during a long and uninterrupted period, resulting to fatigue and, thus, affecting ratings quality.

Concerning the study of the influence of modalities in overall difficulty, it occurred that the movies clips presented with audio only were perceived as the most difficult by the participants. In order to find what may be the principal reasons of this phenomenon, a closer look was taken at the commentaries left by the participants. When they assessed the movie clips with audio only as being difficult the main arguments advanced were: a high speech rate (sometimes leading to mispronunciation), the presence of noisy backgrounds, speakers accents and a lack of context. Most of these factors are related to speech intelligibility: when movie clips with audio only are assessed as difficult it is due to intelligibility issues. Four additional multiple linear regressions were performed, keeping overall difficulty as a dependent variable and lexical and grammatical complexity and speech intelligibility as independent variables. Participants seemed more sensible to speech intelligibility when they only had access to the audio: for A condition, speech intelligibility explained 33% of the overall difficulty, versus 20% for AT condition, 25% for AV condition and 17% for AVT condition. Adding audio and video modalities decreases the mean of overall difficulty and also improves the speech intelligibility (while decreasing its influence on overall difficulty): the easiest movie clips (and also the most intelligible) being the ones where the participants could have access to audio, video and text simultaneously. Considering the previously cited factors leading to high ratings of overall difficulty, the improvements of the comprehension and the intelligibility could be due to the fact that adding text diminishes the effects of high speech rate, noisy background and speakers accents. Concerning contextual issues, video may help to bring additional information, and then to have a better comprehension of the situation. Seeing the speaker faces and particularly lip movements may also help to improve intelligibility, but the contribution of video to listening comprehension should be investigated further. To sum up, improving the speech intelligibility diminishes the perceived overall difficulty and exploiting all the modalities can ensure a high speech intelligibility and a low overall perceived difficulty.

These observations are issued from a subjective study, which can lead to the development of an automatic predictor of overall difficulty and to the obtention of an objective measure. The next step will be to compare the ratings given by the participants with features extracted automatically from the audio modality, the video modality and the text modality. The objective will be first to determine if some of these features can allow to build a model able to predict correct ratings and, then to check if they are consistent with those proposed by this study participants.

6. Acknowledgements

We would like to thank the Association Nationale Recherche Technologie (ANRT) for the funding of the industrial PhD, through which this research work has been carried out.

7. Bibliography

Adank, P., Evans, B., Stuart-Smith, J., and Scott, S. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human perception and performance*, 35(2):520.

Anderson, J. R. (2005). *Cognitive Psychology and its Implications*. Macmillan, New York, USA, 7th edition.

Aneiro, S. (1989). *The Influence of Receiver Apprehension in Foreign Language Learners on Listening Comprehension among Puerto Rican College Students*. Unpublished doctoral dissertation, New York University.

Batel, E. (2014). The effectivness of video vs. written text in English comprehension and acquisition of ESL students. *Arab World English Journal*, 5(4):326–335.

Blau, E. (1991). More on comprehension input: The effect of pauses and hesitation markers on listening comprehension. *Annual Meeting of the Puerto Rico Teachers of English to Speakers of Other Languages*.

Boyle, J. (1984). Factors affecting listening comprehension. *ELT Journal*, 38(1):34–38.

Buck, G. (2001). *Assessing Listening*. Cambridge University Press, Cambridge, England.

Carrow-Woolfolk, E. (1999). Comprehensive assessment of spoken language. *Bloomington, MN: Pearson Assessment*.

Chang, A. C.-S. and Read, J. (2008). Reducing listening text anxiety through various forms of listening support. *TESL-EJ*, 12(1):1–25.

Collins-Thompson, K. and Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462.

Corley, M. and Hartsuiker, R. (2003). Hesitation in speech can... um... help a listener understand. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 25(25):276–281.

Dahl, T. and Ludvigsen, S. (2014). How I see what you're saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, 98(3):813–833.

Dale, E. and Chall, J. S. (1948). A formula for predicting readability: Instructions. *Educational research bulletin*, 27(2):37–54.

Daneman, M. and Merikle, P. (1996). Working memory and language comprehension: A meta-analysis. *Psychonomic Bulletin and Review*, 3:422–433.

Elkhafaifi, H. (2005). Listening comprehension and anxiety in the Arabic language classroom. *The Modern Language Journal*, 89:206–220.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Florit, E., Roch, M., and Levorato, M. (2013). The relation between listening comprehension of text and sentences in preschoolers: Specific or mediated by lower and higher level components? *Applied Psycholinguistics*, 34:395–415.

François, T. and Fairon, C. (2012). An AI readability formula for french as a foreign language. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 466–477.

François, T. (2009). Modèles statistiques pour l'estimation automatique de la difficulté de textes de FLE. *Actes de RECITAL 2009*.

Goh, C. (1994). How much do learners know about the factors that influence their listening comprehension? *Hong Kong Journal of Applied Linguistics*, 4(1):17–42.

Goldin-Meadow, S. and Alibali, M. (2013). Gesture's role in speaking, learning and creating language. *Annual Review of psychology*, 64:257–283.

Gougenheim, G. (1964). *L'élaboration du français fondamental (1er degré): étude sur l'établissement d'un vocabulaire et d'une grammaire de base (Vol. 1)*. Chilton Books, Sudbury, United Kingdom.

Harmer, J. (2007). *The practice of English language teaching*. Pearson Longman, Harlow, United Kingdom.

Hayati, A. and Mohmedi, F. (2011). The effects of films with and without subtitles on listening comprehension of EFL learners. *British Journal of Educational Technology*, 42(1):181–192.

Hayati, A. (2010). The effect of speech rate on listening comprehension of EFL learners. *Creative Education*, 2:107–114.

Henry, G. (1975). *Comment mesurer la lisibilité*. Labor.

Ikeno, A. and Hansen, J. (2006). Perceptual recognition cues in native english accent variation: 'listener accent, perceived accent, and comprehension'. *2006 IEEE International Conference on Acoustic Speech and Signal Processing Proceedings*, 1:401–404.

Jones, L. and Plass, J. (2002). Supporting listening comprehension and vocabulary acquisition with multimodal annotation. *The Modern Language Journal*, 86:546–561.

Kalyuga, S., Chandler, P., and Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology*, 13(4):351–371.

Kellerman, S. (1992). I see what you mean: The role of kinesic behaviour in listening and implications for foreign and second language learning. *Applied linguistics*, 13(3):239–258.

Kim, Y.-S. and Phillips, B. (2014). Cognitive correlates of listening comprehension. *Reading Research Quarterly*, 49:269–281.

Klare, G. (1974). Assessing readability. *Reading Research Quarterly*, 10(1):62–102.

Kruger, J. and Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual text. *Proceedings of Eye Tracking South Africa*, 1:29–31.

Kurita, T. (2012). Issues in second language learning comprehension and the pedagogical implications. *Accent Asia*, 5(1):30–44.

Larsby, B., Hallgren, M., Lyxell, B., and Artinger, S. (2005). Cognitive performance and perceived effort in speech processing tasks: effects of different noise backgrounds in normal-hearing and hearing-impaired sub-

jects. *International Journal of Audiology*, 44(3):131–143.

Lecumberri, M., Cooke, M., and Cutler, A. (2010). Non-native speech perception in adverse conditions: A review. *Speech communication*, 52(11-12):864–886.

Lively, B. and Pressey, S. (1923). A method for measuring the 'vocabulary burden' of textbooks. *Educational Administration and Supervision*, 9:938–398.

Markham, P., Peter, L., and McCarthy, T. (2001). The effects of native language vs target language captions on foreign language students' dvd video comprehension. *Foreign Language Annals*, 34(5):439–445.

Mitterer, H. and J.McQueen. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one*, 4(11):e7785.

Nation, I. (2006). How large vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63:59–82.

Nissan, S., DeVincenzi, F., and Tang, K. (1995). An analysis of factors affecting the difficulty of dialogue items in TOEFL listening comprehension. *ETS Research Report Series*, 2:i–42.

Noro, T. (2006). Developing a construct model of 'listening stress': A qualitative study of the affective domain on the listening process. *Annual Review of English Language Education in Japan*, 17:61–70.

Perez, M., Noortgate, W., and Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, 41(3):720–739.

Perfetti, C. and Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, 18:22–37.

Robinson, B., Mervis, C., and Robinson, B. (2003). The roles of verbal short-term memory and working memory in the acquisition of grammar by children with Williams syndrome. *Developmental Neuropsychology*, 23:13–31.

Sherman, A. L. (1893). *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn & Co.

Stokes, S. and Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *Journal of Child Psychology and Psychiatry*, 50:498–505.

Sueyoshi, A. and Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, 55:661–699.

Tannen, D. (1982). *Spoken and written languages: Exploring orality and literacy (Vol. 32)*. ABLEX Publishing Corporation, New York, USA.

Thorndike, E. L. (1921). *The teacher's word book*. Bureau of Publications, Teachers College, Columbia University, New York, USA.

Traverso, V. (2013). *L'analyse des conversations*. Armand Colin, Paris, France, réédition 2013 edition.

Tunmer, W. (1989). The role of language-related factors in reading disability. *Phonology and reading disability: Solving the reading puzzle*, 6:91–131.

Vandergrift, L. (2007). Recent developments in second and language foreign language listening comprehension research. *Language Teaching*, 40:191–210.

Voss, B. (1979). Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, 22(2):129–144.

Weil, A. (2003). *The impact of perceptual dissimilarity on the perception of foreign accented speech*. Doctoral dissertation, The Ohio State University.

Wong, J. and Waring, H. (2010). *Conversation Analysis and Second Language Pedagogy*. Taylor Francis, Abingdon-on-Thames, United Kingdom.

Yasin, B., Mustafa, F., and Permatasari, R. (2018). How much videos win over audios in listening instruction for EFL learners. *The Turkish Online Journal of Educational Technology*, 17(1):92–100.

Zhao, Y. (1997). The effects of listener's control of speech rate on second language comprehension. *Applied Linguistics*, 18(1):49–68.