



## Audio description text for indexing films

**James M Turner**

École de bibliothéconomie et des sciences de l'information  
Université de Montréal, Montréal,  
Québec, Canada

**Suzanne Mathieu**

Ville de Montréal, Montréal,  
Québec, Canada

Meeting:

**157 Classification and Indexing**

Simultaneous Interpretation:

no

### WORLD LIBRARY AND INFORMATION CONGRESS: 73RD IFLA GENERAL CONFERENCE AND COUNCIL

19-23 August 2007, Durban, South Africa

<http://www.ifla.org/iv/ifla73/index.htm>

#### **Abstract**

*Access to audiovisual materials should be as open and free as access to print-based materials. However, we have not yet achieved such a reality. Methods useful for organising print-based materials do not necessarily work well when applied to audiovisual and multimedia materials. In this project, we studied using audio description text and written descriptions to generate keywords for indexing moving images. We found that such sources are fruitful and helpful. In the second part of the study, we looked at the possibility of automatically translating keywords from audio description text into other languages to use them as indexing. Here again, the results are encouraging.*

#### **Introduction**

As the *IFLA Guidelines for audiovisual and multimedia materials in libraries and other institutions* remind us, access to audiovisual materials should be as open and free as access to print-based materials. However, we have not yet achieved such a reality. Methods useful for organising print-based materials do not necessarily work well when applied to audiovisual and multimedia materials, and much work remains to be done in developing adequate ways to provide good access to users. In this paper, we report on work carried out in the context of the E-Inclusion Research Network (<<http://e-inclusion.crim.ca/?q=en>>). This research network has a goal of “creating powerful audio-video tools... to improve the richness of the multi-media experience for the blind, the deaf, the hard of hearing, and the hard of seeing”. Based at the CRIM (*Centre de recherche informatique de Montréal* or Montréal Centre for Computer Research), the network includes several universities, partners such as the National Film Board of Canada (NFB) and the CNIB/INCA (formerly the Canadian National Institute for the Blind), and AudioVision, a describing agency.

Project 3.1 of the research network involves identification of types of information needed by persons who are blind or with vision loss, in order to understand moving images. The objective of the project was to produce guidelines for persons who describe moving images for this audience. Such persons are called “describers”. In order to study the problem, we viewed a number of films with people who are blind or who have vision loss and asked them to comment on the audio description. We wanted to find out if they found the description adequate, and if not, what improvements they would like to see. One goal of the broader E-Inclusion Research Network is to study the possibility of offering variable description for users, that is to say, description that the users can control, for example by getting more or less description, or having only certain types of information described for them. This is of interest because there is a broad spectrum of vision loss problems, and in fact few users are completely blind. In addition, with aging populations, more and more people are losing vision because they are getting older and living longer. Furthermore, even people who can see perfectly well enjoy watching described versions of films, because information is sometimes given that they would not have noticed otherwise, or that provides some context that helps them understand the film better.

In order to achieve our goal of studying these problems, we needed to identify the characteristics of the films we viewed with people who are blind or with vision loss, and this involved a great deal of background work. In the next section, we describe some of this work. In analysing a number of films, we also studied the question of whether the vocabulary used to describe the image to viewers can also be used as indexing, and that particular problem is the focus of this paper.

## **Context**

Since 1995, we have worked on a number of research projects to study the characteristics of audio description text in order to see how useful it might be in indexing moving images. We are especially interested in the question of indexing individual shots in movies, television programmes, and other audiovisual documents. In studying this question, we developed a methodology for transcribing audio description text and matching it with the image (Turner 1998), then refined the methodology in later projects (Turner and Colinet 2004, Mathieu and Turner 2007). This involves the notion of “episodes” of audio description, or “chunks” of information with some subject unity. When the description changes to some other subject, a new episode begins. Since audio description is provided as an additional sound track and inserted in available sound spaces, it can not always be recited at the same time the image it describes is shown on screen. We studied the patterns for inserting the description, identifying whether it occurred at the same time as the image, one shot before or after the image appeared on screen, two shots before or after, and so on.

By linking the description text in a moving image database to the corresponding image, whether they occur on screen at the same time or not, and by identifying descriptors in the text that can be considered useful indexing terms, we can provide shot by shot indexing to moving images. This is highly desirable for those who make films and television programmes or who study them, and is very helpful in television newsrooms where shot by shot indexing is necessary in order to have material readily available as news stories build from day to day.

In addition, moving image databases that index each shot usually contain a written description of the action in the shot. This helps users skim the text quickly, without actually

viewing the shots, until they narrow the selection down to a short list of shots that can then be viewed and the appropriate ones chosen for use. This approach provides important economies in searching, because the viewing time is greatly reduced.

We wanted to see if there was enough redundancy in the audio description text to use it as a substitute for such descriptions. If so, further economies could be realised in avoiding having to formulate and write the descriptions, and the indexing could possibly be derived exclusively from the audio description text. However, the patterns found in our previous work led us to the conclusion that the audio description text often refers to a sequence of shots, and not necessarily to individual shots. Everything depends on the particular film, but the following examples will hopefully clarify what we mean here. In some films, the action is fast paced, and the editing helps maintain the viewer's excitement by cutting many short shots together. In such a case, it is not necessarily useful to describe each shot, but rather what is happening in the sequence as a whole. In other films, the action is slow, individual shots last a long time on the screen, and there is little competition for the sound spaces because there may be only background music and no dialogues. In such a case, the describer has ample time to talk about the content of the image. And in some of the films we studied in the context of the present project, such as the experimental animation films of Normal McLaren, the films are short, and have only music for sound. Furthermore, we can say that the whole film is a single shot, or not even a shot as such, in cases where the image is scratched or drawn directly onto the film stock. Thus films and other audiovisual products come in a wide variety of forms, and we can not formulate overall rules that fit all of them. Rather, we need to work out the patterns for various types of products.

### **The E-Inclusion Network projet**

This research addresses three aspects of the programme theme of the Classification and Indexing Section: minority language subject access, cooperating with users in designing subject access tools, and facilitating and extending subject access to library resources. Minority language subject access is addressed through the bilingual (English and French) aspects of our project, and although we do not discuss the issue of multilingual access in this paper, the work it reports can be tied to some of our previous work (for example Turner and Hudon 2002) to extend the coverage to other languages. Cooperating with users takes place through our discussions with people who are blind or with low vision, who gave us a great deal of feedback on what aspects of the audio description text they found useful in understanding the films we screened together. In analysing the topics discussed during these sessions, we were able to formulate our guidelines. In addition, their comments on the types of description they find helpful offer guidance as to what the most useful kinds of indexing terms are for this audience. Facilitating and extending subject access is addressed by providing any indexing at all at the sequence and shot level, as these are presently unavailable, except in specialised situations such as stockshot libraries and television newsrooms.

In our analysis of the audio description text of the films we studied, we sought to validate the typology of the kinds of description provided in films we studied in our previous projects. By adding to the small corpus of films already studied, we were able to consolidate our findings by comparing them to the results we had already obtained, in order to see whether it was necessary to introduce new categories of description, or whether the typology we had already built included the kinds of information described in the new films we studied. Since

so little work of this kind has been carried out (hence the references mostly to our own work), every new film studied adds to the corpus of films analysed and helps confirm or contradict our results.

In relation to the results we report in this paper, our objectives for the project were :

- to validate the typology of kinds of description offered in audio description text
- to compare the keywords in the audio description text with the keywords in the user descriptions written for each shot, in order to estimate the possibility of deriving indexing from these texts
- to compare the English keywords with the French keywords in order to estimate the possibility of providing automatic indexing between these two languages and in other languages

By meeting these objectives, we hoped to gain enough knowledge to obtain adequate responses to our research questions. In the next section, we discuss the methodology we used to achieve these aims.

## **Methodology**

Within the framework of this project, we analysed eleven films, of which two were feature-length films, two were documentaries, and seven were animated shorts. The analysis consisted of breaking down each film into sequences and individual shots, breaking down the audio description text into episodes, and providing a free text description for each shot. This latter task was carried out by a research assistant who viewed the films without the sound, so as to avoid being influenced by the dialogues, the narration, or the audio description. This work resulted in the master files we used for analysis.

We used available software to match individual segments of the films to time code, as a way of providing precise identification. The time code included hour:minute:second:image. For films provided by the National Film Board of Canada, a DVD version with the time code burned in was made for us.

In order to address our specific research objectives, we categorised the episodes of audio description, using the typology we developed in a previous project (Turner 1998) and refined in a later project (Turner and Colinet 2004). We also considered using for this analysis the typology presented by Piety (2003) but judged it too general for our purposes. The *ITC guidance on standards for audio description* (Independent 2000) helped shed further light on this process.

For each production analysed, we created Excel files to house the data. These include the breakdown of the film into sequences and shots, each with its description and the accompanying audio description text (figure 1). For the feature-length films, each chapter is stored in a tab, This method helped us get an overview of the data, as well as to automate some of the calculations which gave us our results. This represents a refinement of the techniques we used in our previous work, which used database software that houses

information for individual shots but that does not provide an overview of the data.

The Excel files thus contain numerous data elements and permit various kinds of analysis. In this paper, we focus on aspects related to the Classification and Indexing Section, first on the relationship between the audio description text and the written description, and second, on

#	Chapitre	TC in	TC out	Type	AV	Description	Texte de l'audiovision	
7	7	00:45:35	00:45:36	Décor?	1	Gros plan st	Photos de l'homme chauve que Nino poursuivait.	
8	7	00:46:26	00:46:26	Action	1	Plongée sur	Recouvre la toile.	
9	7	00:47:25	00:47:28	Action	2	Amélie est d	Dans son lit en fer forgé, Amélie regarde à nouveau les photos.	
10	7	00:47:33	00:47:33	Action	2	Amélie, ass	Les visages se succèdent.	
11	7	00:47:36	00:47:36	Person	2	Photos d'un	Un homme barbu.	
12	7	00:47:38	00:47:38	Person	2	Photos d'un	Une femme qui éternue.	
13	7	00:47:41	00:47:43	Person	2	Photos d'un	Un jeune homme noir à l'air sérieux.	
14	7	00:47:45	00:47:45	Person	2	Photos d'un	Un bébé	
15	7	00:47:46	00:47:47	Person	2	Photo d'un l	Un homme qui remet son bonnet.	
16	7	00:47:50	00:47:52	Action	2	Amélie est c	Amélie est endormie, la télécommande de son téléviseur posée près d'elle.	
17	11	7	00:47:54	00:47:56	Action	2	On voit une	La télévision diffuse une course cycliste.
18	12	7	00:47:56	00:48:00	Action	2	On voit la c	Un cheval s'échappe de son pré en sautant une barrière et se met à galoper au milieu des coureurs.
19	13	7	00:48:01	00:48:03	Action	2	Amélie se r	Amélie ouvre les yeux, enregistre la séquence.
20	14	7	00:48:12	00:48:14	Action	1	On voit les	Amélie descend les escaliers de son immeuble.
21	15	7	00:48:16	00:48:19	Action	1	On voit Amé	Elle s'arrête sur un palier et dépose une grosse enveloppe craft sous un paillason.
22	16	7	00:48:21	00:48:24	Action	1	Amélie se r	En se redressant, elle voit un trousseau de clefs oublié sur la porte d'en face.
23	17	7	00:48:27	00:48:29	Action	1	Amélie pren	Elle le prend et sort.
24	18	7	00:48:30	00:48:32	Action	1	Amélie desc	Dehors il y a du brouillard, l'épiciier derrière son étal.
25	19	7	00:48:41	00:48:42	Action	1	Lucien tourn	Il se tourne vers son commis.
26	20	7	00:48:53	00:48:53	Action	1	L'épiciier tap	Il le tape.
27	21	7	00:49:03	00:49:04	Action	1	Lucien baiss	Lucien s'éclipse, Amélie est furieuse.
28	22	7	00:49:21	00:49:23	Action	1	Amélie man	Elle marche dans la rue d'un pas décidé.
29	23	7	00:49:24	00:49:27	Action	1	Amélie man	Pousse la porte du coordonnier sur laquelle est inscrit: Ici clé-minute.
30	24	7	00:49:31	00:49:33	Action	1	On voit Amé	Sur le palier, elle remet le trousseau de clefs à sa place.
31	25	7	00:49:34	00:49:34	Décor?	1	Amélie se d	Dans sa poche, une clef apparaît comme par radioscopie.
32	26	7	00:49:39	00:49:39	Action	1	Amélie rede	Elle redescend.
33	27	7	00:49:40	00:49:41	Action	2	Amélie rede	Georgette est au kiosque à journeaux.
34	28	7	00:50:00	00:50:02	Action	2	La dame du	Elle tend une pile de journeaux à Georgette qui semble ne pas comprendre.
35	29	7	00:50:03	00:50:04	Lieu	1	Gros plan st	Au café, Hipolito est au comptoir.
36	30	7	00:50:46	00:50:48	Action	1	Gina remplac	Elle remet en place le cou d'un client.
37	31	7	00:50:59	00:51:01	Action	1	On voit José	Georgette et Joseph se regardent du coin de l'œil.
38	32	7	00:51:09	00:51:11	Action	2	On voit Luci	Voulant regarder sa montre, Lucien renverse le cageau.
39	33	7	00:51:13	00:51:15	Action	2	Amélie lanci	Amélie regarde fixement l'épiciier.
40	34	7	00:51:17	00:51:19	Action	1	Amélie débi	Avec sa clef secrète, elle entre chez Collignon.
41	35	7	00:51:23	00:51:24	Action	1	Amélie man	Elle traverse le couloir.
42	36	7	00:51:27	00:51:29	Action	1	Se tourne vi	Passe dans le salon où une paire de charentaises attend au pied d'un fauteuil.
43	37	7	00:51:33	00:51:36	Action	1	Amélie soul	Elle soulève les pantoufles pour regarder la peinture.
44	38	7	00:51:37	00:51:37	Image	1	On voit le c	42
45	39	7	00:51:39	00:51:41	Action	1	Amélie déto	Elle ouvre le placard à chaussures.
46	40	7	00:51:46	00:51:46	Action	1	Amélie rega	Lime des lacets.
47	41	7	00:51:56	00:52:00	Action	1	Amélie se d	Dans la salle de bain, elle intégrant le dentifrice avec un tube de crème pour les pieds

the relationship between keywords in two languages, English and French.

Figure 1. An Excel sheet with audio description text matched to the shots.

For each shot, by comparing the text of the written description with that of the audio description, we can estimate the frequency with which the same keywords occur in both texts. This gives some indication as to the possibility of automating the process of indexing, as well as to which of the two kinds of text is most useful for this purpose. We made a copy of the spreadsheet for the film *Le fabuleux destin d'Amélie Poulain*, one of the films that was studied in a various ways by a number of researchers in the broader context of the E-Inclusion project. This version of the spreadsheet (figure 2) also shows Chapter 7 of the film, re-arranged to show only columns A, G, and H, respectively the shot number, the written description, and the audio description text. Keywords that occur in both the written description and the audio description text are shown in orange. Note that the corresponding keywords do not always fit the same shots, for example in shots 22-25. This is because the audio description is inserted into the film in available sound spaces. For this analysis, we have thus allowed including keywords that occur within a few shots of the one they are meant to describe.



In an analysis of six chapters of this film, there are a total of 238 shots. Of these, exactly 50 have no keywords in the written description that also appear in the audio description. For the remaining 188 shots, there are 463 keywords that appear in both descriptions. The ratio, then, is approximately 2.5 keywords per shot in common.

In answer to our question of whether the written description is more useful than the audio description in producing keywords that could be considered useful for indexing, we observe that the essential descriptors appear in both texts, as one might expect. However, in each text there are a number of additional keywords (left in black) that do not appear in the other text. We can conclude that it is fruitful to use both texts for generating keywords that correspond to the image, insofar as the process can be automated. If we need to choose only one of the two, the correspondence between the essential descriptors is such that either text could be used to generate useful indexing.

To compare keywords in English and French, we used a short film entitled *Home Security / En toute sécurité*, produced by the National Film Board of Canada, and that had audio description available in both languages. We put each keyword from the original English into the dictionary, and looked to see if the French translation supplied by the dictionary contained the French word we recorded from the audio description script. We created an Excel spreadsheet to analyse the data in three columns (figure 3). Column A lists the keywords in English, as taken from the script for the audio description, supplied by the National Film Board. Column B lists the corresponding keywords in French, taken from the French-language audio description script provided by the National Film Board. Column C shows the results from Reverso Online / Collins dictionaries. Column D contains an X where there is a match, and Column E contains an X where there is no match.

	A	B	C	D	E
1	Keyword	Mot clé	Reverso	OUI	NON
14	ball	balle	balle	X	
15	bark	aboyer	aboyer	X	
16	barred	recouvertes de barreaux	barricadé		X
17	baseball	baseball	base-ball	X	
18	bat	batte	batte	X	
19	beaten	reçoit un coup	battu		X
20	beats	cogne	battre		X
21	behind	derrière	derrière	X	
22	bells	sonneries	sonnerie	X	
23	biker	motard	motard	X	
24	blue	bleu	bleu	X	
25	board	office	conseil		X
26	bolts	part en courant	filer		X
27	bombs	bombes	bombes	X	
28	boys	garçons	garçons	X	
29	breaking into	entre	s'introduire par effraction		X
30	breaks	se brise	se brise	X	
31	building		0	0	
32	bulges		0	0	
33	burglar	voleur	cambrioleur		X
34	butterfly	papillon	papillon	X	
35	button	bouton	bouton	X	
36	cage	cage	cage	X	
37	Canada	Canada	Canada	X	
38	car		0	0	
39	cat	chat	chat	X	
40	ceiling	chandelier	plafond		X
41	channel	chaîne	chaîne	X	
42	chase	poursuit	poursuit	X	
43	chattering	claquant	bavarder		X
44	city	ville	ville	X	
45	closes	fonce	fermer		X
46	clothes	vêtements	vêtements	X	
47	commercial	publicité	publicitaire	X	
48	content	content	content	X	
49	continue	continuent	continuent	X	
50	cops	policiers	flics		X
51	crashes	traverse	percuter		X
52	criminal	bandit	criminel		X

Figure 3. The English keywords with French equivalents, and keywords generated by the automatic translator..

In this five-minute film, there were a total of 252 keywords in English in the audio description script. Of these, 43 had no equivalent in French (because of linguistic constructions in the way the French-language version was made). These are shown in the spreadsheet with a zero in columns B and C. Of the remaining 209 terms, 154 had an equivalent in French and 55 did not. Thus approximately 74% of the terms had a direct equivalent. This result is rather encouraging in the context of using such keywords to produce indexing, because it means that 74% of the time, an automatic translator will generate useful keywords in another language (at least between English and French). However, since synonyms were not taken into account in the analysis, if the translation software were tied to an online thesaurus, the score would be higher. In the example we give in Figure 3, we note that in most of the cases where there is not a match, the terms in columns B and C can be considered synonyms or closely-related terms.

## Discussion

The results reported here indicate that text created for other purposes can point to individual shots or sequences of a film and thus be used as a way to index these shots or sequences. In the case of keywords from written descriptions or from audio description text, enough useful keywords are found to index shots or sequences. Combined with keywords from subtitles created for the deaf and hard of hearing, which cover the dialogues heard in the films, fairly complete indexing can be provided at the shot and sequence levels. Applying these research results involves finding practical ways to exploit the text files for keywords and for attaching the keywords to the shots and sequences. We imagine this as occurring

rather naturally over time, as production becomes more and more digital, and as large databases for managing the digital assets from a movie are created. In other words, text modules created as post-production activities will already form part of the production database for the film, so it is just a question of programming the automatic generation of keywords, then automatically linking them to shots and sequences. These activities are relatively easy to programme. Many more possibilities for exploiting such text can be imagined, such as linking them with vocabulary management tools to fit them into thesaural structures or to generate indexing in other languages.

The animated films we studied in the context of this project are interesting because they exhibit new patterns that did not occur in the films we studied in previous projects. There are a number of reasons for this. Some of the films were experimental, including techniques such as drawing or scratching the image directly on the film stock, so that they do not include actual shots, since the films were not made using a camera. Other films used pixilation, a stop-motion technique with actors moving a little between still shots to give an animated effect. Here again, the notion of “shots” does not apply.

In addition to the questions dealing with the notion of shots, the animated films were often without dialogues, so that there were no competing sound spaces for the audio description. This resulted in the episodes of audio description being recited almost always (98.29% of the time) at the same time that the image they described was shown on screen, instead of before or after, as is often the case with live action movies. Thus, in some of the films we studied, we found patterns different from those found in our previous work.

Another issue that became clear in the course of this work was that the user base of visually-impaired people includes such a variety of vision loss problems and cognitive styles among users that it would be highly desirable to offer to users the possibility of adjusting the audio description for any film they are watching. Various levels of detail in the amount of description would be one aspect. In preparing the audio description text, episodes could be tagged as Level 1, 2, or 3, for example, and the user could then indicate which level they wished to hear while the film is playing. Another aspect involves the kind of things being described. Our typology could be used to offer various possibilities to users, and those who can see enough to know whether it is sunny or cloudy, for example, could then choose not to hear this information. These examples illustrate the kind of problems we hope to study in the next phase of the project.

## **Conclusions**

This paper reports on two aspects of the work we did in the context of the E-Inclusion Research Network project, and other papers will report on other aspects. Here we looked at the usefulness of recycling keywords found in text produced for the purpose of describing the image to persons who are blind or who have some vision loss, as indexing to individual shots and sequences of films. We found that such keywords constitute helpful indexing terms often enough to use them as such.

The other aspect we considered is whether keywords can be generated automatically in another language, in order to produce bilingual and ultimately multilingual indexing. By measuring simple correspondence between English and French terms, not accounting for synonyms or near-synonyms, we found that there is correspondence three-quarters of the

time, easily enough to make such a practice useful.

This study contributes to work on the more general problem of using available text created for other purposes as a source for mining indexing terms that describe individual shots or sequences of a movie or a television programme. Together with a number of other kinds of such text, the kinds we studied here form a rich corpus of available keywords that can be recycled as indexing terms insofar as the processes for producing them can be automated. Thus access to individual shots and sequences, in general far too costly to produce by human indexers, can be achieved.

## Acknowledgements

We thank Canadian Heritage for its contribution to funding the E-Inclusion Research Network. We thank our partners at the Centre de recherche en informatique de Montréal (CRIM), for their help and guidance and for providing a collegial atmosphere in which to work. We are very grateful to Anne Jarry at the Canadian National Institute for the Blind (CNIB) for her help with organising viewing sessions with blind and partially-sighted participants, and for her invaluable insights and counselling. We are grateful to all the participants who agreed to provide us with data by offering their views and opinions on how to make audio description more useful. Our thanks to Alain Lapierre, also of the CNIB, for organisational and technical assistance. Finally, we would like to thank Linda Rivest, our research assistant, who prepared much of the data for analysis.

## References

- Independent Television Commission [UK]. 2000. ITC Guidance on standards for audio description / Nick Lodge.  
<[http://www.ofcom.org.uk/tv/ifi/guidance/tv\\_access\\_serv/archive/audio\\_description\\_stnds/itcguide\\_sds\\_audio\\_desc\\_word.doc](http://www.ofcom.org.uk/tv/ifi/guidance/tv_access_serv/archive/audio_description_stnds/itcguide_sds_audio_desc_word.doc)>.
- Mathieu, Suzanne and James M Turner. 2007. Audiovision ou comment faire voir l'information par les personnes aveugles et malvoyantes : lignes directrices pour la description d'images en mouvement. Canadian Association for Information Science, annual conference, *Proceedings* [in press].
- Piety, Philip J. 2003. Audio description, a visual assistive discourse: an investigation into language used to provide the visually disabled access to information in electronic texts. Master of Arts thesis. Washington: Georgetown University.
- Reverso Online. 2007. <<http://www.reverso.net>>.
- Turner, James M. and Emmanuël Colinet. 2004. Using audio description for indexing moving images. *Knowledge Organization* 31, no. 4, 222-230.
- Turner, James M. and Michèle Hudon. 2002. Multilingual metadata for moving image databases: preliminary results. *Advancing Knowledge: Expanding Horizons for Information Science, Proceedings of the 30th annual conference of the Canadian Association for Information Science*, ed. Lynne C. Howarth, Christopher Cronin, Anna T.

Slawek. Toronto: Faculty of Information Studies, 34-45.

Turner, James M. 1998. Some characteristics of audio description and the corresponding moving image. *Proceedings of the 61st ASIS annual meeting, Pittsburgh, PA, October 24-29, 1998*, ed. Cecilia M Preston. Medford, NJ: Information Today, 108-117.