



## **Gaumont Selects SYSTRAN to Translate its French Online Newsreel Catalogue for the Global Marketplace**

**By,  
Andrew Joscelyne**

October 2002

When Stuart McKay, a freelance film archive researcher, was looking for historical footage of the First World War for a new British TV series, he naturally turned to the Web for information about relevant holdings in French film archives. France, the birthplace of cinema, has a vast range of holdings of early film in its military and government archives. But what made Stuart's search particularly fruitful was the availability of a remarkable online database listing the complete historical footage held in the Cinématique Gaumont newsreel archives. And above all, the ability to search this database in the universal language of professional media searchers – English.

The Cinématique Gaumont is a French film library offering the largest range of French language newsreel and other film holdings of its kind. Information about the archive is accessible via the Internet, offering film researchers an unparalleled database for searching cultural and historical material on celluloid, viewing excerpts and then ordering them. To render this facility as universally accessible as possible to researchers like Stuart McKay, Gaumont contracted SYSTRAN to provide an on-demand translation solution that would enable researchers to retrieve the database and read its film descriptions in English. Gaumont has been able to substantially grow its market for newsreel film archive users by combining the ease of access afforded by an online database with the communicative effectiveness of SYSTRAN's Machine Translation technology.



Headquarters  
SYSTRAN S.A.  
1, rue du Cimetière - BP7  
95230 Soisy-sous-Montmorency  
France  
Tél. : + 33(0)1 39 34 97 97  
Fax : + 33(0)1 39 89 49 34

North America  
SYSTRAN Software Inc.  
9333 Genesee Avenue  
Plaza Level, Suite PL1  
San Diego, CA 92121 - USA  
Tel. : + 1 858 457 1900  
Fax : + 1 858 457 0648



### ***Leveraging Gaumont's Film Assets***

The French company Gaumont began life as **la Société des Etablissements Gaumont** in 1906 with an ambition to capture real life on film. It is therefore the longest lasting of any of the film companies created in the heroic years of the film industry. By combining its own holdings with those of another newsreel company, Eclair, Gaumont has one of the richest collections of French newsreel content on record, covering every aspect of French political, social, cultural and sports history from 1895 to 1980.

The company's entire visual treasure trove contains some 90,000 reels of Gaumont newsreels, as well as over 10,000 km (6,000 miles) of French film stock, and 6,000 hours of fiction films, as well as a vast collection of 10,000 still photographs. The vital newsreel material ranges from colored footage using a process developed by Léon Gaumont of the 1918 victory parade at the end of WWI to the Tour de France in 1980.

Today, these film assets contribute greatly towards enriching media products such as documentary and advertising films, multimedia products, museum presentations and online cultural events. As such they create a vital source of new business. A sequence of newsreel lasting just a few seconds can cost several thousand dollars, thus establishing the global market for newsreel film assets a multimillion-dollar business.

Due to the linguistic nature of its catalogue data, Gaumont's major market has until recently been France and French speaking cultures from Quebec to Switzerland. Traditionally, producers in search of material would have to spend time traveling to an archive site and searching through the records to find relevant content. The arrival of the Web has now created an ideal platform for trading between searchers and film asset providers, making it imperative for Gaumont to make its catalogue available online to media buyers in the most accessible idiom – global English.

Headquarters  
SYSTRAN S.A.  
1, rue du Cimetière - BP7  
95230 Soisy-sous-Montmorency  
France  
Tél. : + 33(0)1 39 34 97 97  
Fax : + 33(0)1 39 89 49 34

North America  
SYSTRAN Software Inc.  
9333 Genesee Avenue  
Plaza Level, Suite PL1  
San Diego, CA 92121 - USA  
Tel. : + 1 858 457 1900  
Fax : + 1 858 457 0648



### ***The Gaumont Database***

In 1987, Gaumont took the momentous decision to digitize its catalogue of newsreel products in the form of a database. With three people working full time, the task took ten years, and involved the scanning in and checking of some 120,000 text articles originally kept in notebooks describing the content of each film or newsreel sequence. As the Gaumont Cinématique Manager Manuela Padoan in charge of the project recalls, "in some case this meant actually inputting the content of handwritten entries over 100 years old."

Once the catalogue database was completed, Gaumont then began to digitize a considerable number of the actual newsreels so that media researchers could visualize the contents of a brief sequence before deciding whether to purchase. This in turn amplified the appeal of the service for potential buyers.

Then came the vital question: what is the fastest and most cost-effective way to translate the collection of catalogue articles – containing a total of 14 million words or 104 million characters - into English, thereby opening the holding to the global media market?

### ***Figuring Out Translation***

The catalogue articles describing each newsreel vary in length and content, of course, but they include a large quantity of repeated words and abbreviations, a vast number of proper names of statesmen, politicians, sports people, actors, and people in the news, that naturally were not to be translated, together with a number of 'index' phrases without context that indicated the general topic of the film sequence.

Gaumont calculated that if they had asked a translation team to work on this entire base of articles, the job would have taken at least 10 person years, and demanded extensive coordination and quality management requirements. Moreover, given the

Headquarters  
SYSTRAN S.A.  
1, rue du Cimetière - BP7  
95230 Soisy-sous-Montmorency  
France  
Tél. : + 33(0)1 39 34 97 97  
Fax : + 33(0)1 39 89 49 34

North America  
SYSTRAN Software Inc.  
9333 Genesee Avenue  
Plaza Level, Suite PL1  
San Diego, CA 92121 - USA  
Tel. : + 1 858 457 1900  
Fax : + 1 858 457 0648



summary and repetitive nature of the descriptions, consistent quality over 120,000 documents would have been particularly demanding for human translators.

To complicate matters even further, Gaumont's staff are constantly updating the articles themselves, adding further information about the content of each newsreel sequence after visualization to increase the relevance of the database for searchers. For example, a newsreel originally made to cover a fashion show in 1935 turned out to include as-yet unnoticed footage of the famous French actress Michelle Morgan. The chosen translation solution therefore had to be flexible enough to handle this constantly updated information, in addition to the original content.

Given the volume of translation, the fact that the source documents were all in digital format, and the need to avail the translated version available for information searching purposes rather than for elegant reading as soon as possible, Gaumont opted for a machine translation solution that exploited the power of the computer to provide an adequate response to these needs. They contacted SYSTRAN, who were able to deliver a complete translation of the 14 million words in just nine months.

### ***The SYSTRAN Solution***

SYSTRAN set up a seven-person team of computational linguists and software engineers that also included a project manager, who began by examining the Gaumont corpus of French texts, analyzing the quality and evaluating the work that would be needed to transform the raw French material into a machine tractable source text. They then worked with the Gaumont team – the project manager, archivist, network administrator and translator – to test their developing translation solution and fine-tune it to Gaumont's specific needs.

First the computer files containing the database were enhanced by the removal of typographical errors, spelling variants for the same expressions, and irrelevant markers. A standardized approach was adapted to displaying dates, French cultural terms and especially family names to avoid unwanted English translations of certain

Headquarters  
SYSTRAN S.A.  
1, rue du Cimetière - BP7  
95230 Soisy-sous-Montmorency  
France  
Tél. : + 33(0)1 39 34 97 97  
Fax : + 33(0)1 39 89 49 34

North America  
SYSTRAN Software Inc.  
9333 Genesee Avenue  
Plaza Level, Suite PL1  
San Diego, CA 92121 - USA  
Tel. : + 1 858 457 1900  
Fax : + 1 858 457 0648



French proper nouns. All the information on the quality of the source texts were fed back to the Gaumont team. For example, around 100,000 spelling errors were corrected in the source text.

From the linguistic point of view, metrics were developed to measure the occurrence of potentially ambiguous French terms, phrases and idioms, and the exact nature of the short, verbless sentences that occur in many of the articles. At the same time, the SYSTRAN team collaborated with the Gaumont archive team and translator on developing a series of additional dictionaries to augment SYSTRAN's standard French-English dictionaries. These specific dictionaries (with over 110,000 entries) cover such word types as Patronyms, Names, Toponyms, Cinema terms, and Film Titles.

When all of these resources had been developed and applied to the automatic translation process, they delivered 80% comprehension during a first evaluation. Subsequent work was devoted to enriching the dictionaries and fine-tuning the specific tools required to improve translation quality to Gaumont's desired standard.

***Example of the Machine Translation of a Typical Newsreel Article:***

Note that the input error - 'pours' instead of 'pour' - in the French text is left untranslated in the English version, since 'pours' is already a meaningful word in English, though obviously not in this context.

**French**

A PARIS. Rendez-vous avec **Maurice CHEVALIER**. Pours ses rendez-vous, **Maurice CHEVALIER** a été éclectique. C'est ainsi que la danse, en la personne du marquis de CUEVAS, y cotoie les plus grands noms de l'écran. Michele MORGAN s'y rencontre avec les mannequins de la haute couture qui, eux mêmes, ont pour plus proche voisine Martine CAROL, tandis qu'au générique on relève les noms des plus grandes vedettes du théâtre, du cinéma et de la chanson, Edith PIAF, Jean-

Headquarters  
SYSTRAN S.A.  
1, rue du Cimetière - BP7  
95230 Soisy-sous-Montmorency  
France  
Tél. : + 33(0)1 39 34 97 97  
Fax : + 33(0)1 39 89 49 34

North America  
SYSTRAN Software Inc.  
9333 Genesee Avenue  
Plaza Level, Suite PL1  
San Diego, CA 92121 - USA  
Tel. : + 1 858 457 1900  
Fax : + 1 858 457 0648



Claude PASCAL, Gilbert BECAUD, Eddie CONSTANTINE. **Maurice CHEVALIER** a voulu que tout PARIS participe à sa première rencontre avec la télévision. Chanteur, chanteuse, chanson, acteur, actrice, comédien, comédienne, émission de télévision.

### English

IN PARIS. RENDEZ-VOUS WITH **MAURICE CHEVALIER**. Pours its rendez-vous, **Maurice CHEVALIER** were eclectic. Thus the dancing, in the person of the CUEVAS Marquis, borders on the greatest names of the screen. Michele MORGAN meets with the mannequins of the Haute couture which, themselves, have for more near are on friendly terms Martine CAROL, while to the credit titles the names of the largest stars of the theatre, of the cinema and of the song, Edith PIAF, Jean Claude PASCAL, Gilbert BECAUD, Eddie CONSTANTINE are raised. **Maurice CHEVALIER** wanted Tout Paris to take part in its initial meeting with television. Singer, singer, actor, actress, actor, actress.

### *About the Author*

**Andrew Joscelyne** is a Paris-based language industry consultant and writer who has tracked translation and other language technology solutions for a number of publications. He has been editor of *Language Industry Monitor* and *LE Journal*, and works on awareness-raising projects in language technology issues for the European Commission.