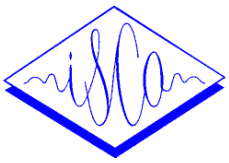

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

MEDIATEX-TASF: A CLOSED CAPTIONING REAL-TIME SERVICE IN FRENCH

Raymond Descout(*), Robert Bergeron(*),
Bernard Mérialdo(**)

(*) Canadian Workplace Automation Research Centre,
CWARC - Multimedia Systems
1575 Chomedey Blvd., Laval (Quebec) H7V 2X2 Canada

(**) IBM-France Scientific Center, Language Industry Dept.,
54 Roger Salengro Blvd., 94126 - Fontenay-sous-Bois France

ABSTRACT

MEDIATEX-TASF is a recent development undertaken for Canadian Broadcasting Corporation's French network by a joint effort from Canadian Workplace Automation Research Centre and IBM-France Scientific Center. This system generates real-time captions during live broadcast by using an experienced stenotypist who is connected to an automatic computer-based transcription system. Real-time closed captioning for hearing-impaired television viewers has been available in English for 10 years, but until now, none were for French-speaking users.

MEDIATEX-TASF adapts the core technology developed by IBM-France (TASF) to the numerous constraints of live operation for providing high quality captioning. Three major improvements were carried out: adaptation of Grandjean method, introduction of specific dictionaries containing more than 6,000 expressions to complement the basic TASF lexicon and post-processing by 500 rules solving lexical problems as well as verb conjugation and word agreement in gender and number.

The error rate of MEDIATEX-TASF is now reaching 5%, which could be sufficient for a commercial usage. First on-the-air tests will occur next September in Montreal.

10.21437/ICSLP.1992-402

1. INTRODUCTION

In November 1988, CBC's French network expressed CWARC a need for improving the quality of their service to hearing-impaired viewers. They expected to have a practical solution for converting real-time speech into written captions appearing at the bottom of the screen.

After reading so many papers related to the advances of speech recognition systems, it was obvious to them that this technology could solve their problem. Unfortunately, voice-recognition technology is still a long way off from being a practical solution, since it would have to recognize words spoken by all speakers at varying rates and pronunciations, deal with newly coined words, distinguish one speaker from another when two people are talking at once, and distinguish speech from background noises.

Such complex cognitive process will remain far from our capacity before any drastic changes in knowledge - as well as in computer processes - will arrive. I'm pretty sure that we won't ever be able to accomplish such an outstanding achievement before a very long time.

However, getting today a working system could be reachable at two conditions: a) replace the well known complex acoustic-to-phonetic level (found in every speech recognition system) by a human being transcribing oral language into a special code by the means of stenotyping machine; b) restrict, at a given time, the application to a semantic domain which is known in advance by the user. In English such a system, called CAT (Computer-Aided Transcription), is used for 10 years. In French two experimental systems were proposed for more than 5 years: one by LIMSI [1] and the other one by IBM-France [2,3]. However, at this time these systems remained still far from a real-time commercial service.

2. REAL-TIME CAPTIONING FOR HEARING-IMPAIRED

Captions or subtitles are transcriptions of audio dialogues and commentaries that are displayed, usually at the bottom of the television screen, for deaf or hard of hearing viewers. Captions can also have an educational role (reading) or can be used for multilingual access to television programs [6].

Captioning can be *open* that is, visible by all viewers, or *closed* or hidden except to concern audiences. Closed captioning works by encoding digital information on line 21 of the vertical blanking interval (VBI) of the television signal. A decoder attached (or built into) a television set processes the information and, by means of an internal character generator, superimposes the captions on the television screen.

Captions have to be prepared ahead of time: a few weeks or days in the case of films or prerecorded broadcasts, or at least a few hours before airing for scripted live programs.

In situations where there is no prior text or script available such as conferences, live interviews, debates, question periods or live broadcast (special news bulletins, a presidential address, crises, Olympic Games, etc.), captioning has to be added in real-time, by synchronization with the audio portion of the program.

In the spring of 1982, the National Captioning Institute (NCI) offered deaf viewers portions of the Academy Awards presentation broadcast for the first time. Today, more than 70 hours of nationally broadcast programming are captioned weekly with this technology. In addition to broadcast applications, CAT is usually used at the court for providing the mandatory *verbatim* of the debates. Today more than 16,000 court reporters (greater than half of the profession) use it in their work.

3. STENOTYPY

For generate real-time captions, typing in words letter by letter on a standard keyboard will not do (a maximum of 100 words per minute can be achieved during a short period of time). It is extremely difficult to keep up with fast, or even moderately paced, speakers for extended periods of time (which can speak up to 220 words per minute) or to ensure accuracy of spelling, particularly names of people and places.

Real-time systems used for TV broadcast in English are an extension of a technology originally developed for court reporters. They do not type in the words letter by letter, but rather make use of shorthand codes. Each stroke may correspond to a syllable, a full word, or even a phrase. The stenotypist is able to keep up with speakers because there are fewer keystrokes per word than are necessary on a

standard computer keyboard. For example, on a steno machine, the word "institute" may be written in only three strokes, i.e., *INS TI TUT*.

Over 25 years ago, work began on developing a computer system that could recognize the shorthand codes and convert them into regular English words, thus eliminating the transcribing task of a court reporter and making it possible to produce finished transcripts quickly. Early systems achieved good-quality translation, but did not work quickly enough to display the English text immediately after the words were spoken. In addition to that, till now, French and the other languages, were reluctant to that technology.

For English, the computer is loaded with a set of dictionaries containing entries for each shorthand stroke or stroke combinations. Each such entry is "defined" as the desired English equivalent. When the stenotypist enters these strokes in the correct order, the computer matches the sequence with the English "definition" and outputs the English.

In French, the stenotypy method is quite different. This method, originally created by Grandjean [4], is more standardized than the English one. It allows a stenotypist to read, and transcript into text, the codes taken by another operator. This is not the case in English where each "definition" is specifically entered by each user. Thus, for the English method, each stenotypist have to customized its own CAT environment before being able to work. Because of this "inter-reading" process, TASF algorithm is well suited for serving a large set of operators trained by Grandjean method.

Grandjean method uses a special 21 keys keyboard as shown in Fig. 1.

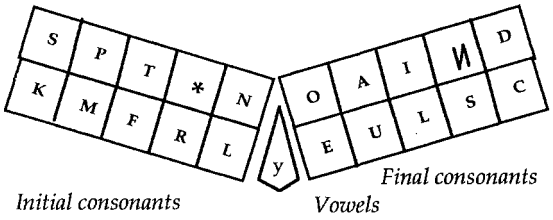

Fig. 1 - Grandjean Stenotyping Keyboard

The method is based on a pseudo-phonetic coding of speech, syllable by syllable. A syllable is typed by striking several keys at the same time, like playing a chord on a piano. The 10 leftmost keys are used for the initial consonants, the 6 middle keys for the vowels, and the 5 rightmost keys for the final consonants. As there are less keys than phonemes, several phonemes are confused on the same key.

In the standard Grandjean method, the coding was :

- *for initial consonants:*
  S = s or z, K = k or g,
  P = p or b, T = t or d, F = f or v
- *for the vowels:*
  no distinctions between open or closed vowels
  E = é or è, the mute e is omitted,
  nasals are coded with final N
- *for the final consonants:*
  N = m or n, L = l or r, S = f, v, s or z
  D = p, b, t or d, C = k or g.

The stenotypist types on a mechanical keyboard from Grandjean company, in which electromechanical contacts are transmitted to the computer through RS-232 interface by the means of a back-box called "Logistène".

Coded syllables are called stenograms. For the most part, no special stroke is used to indicate the end of a word (comparable to the space bar on a computer keyboard). Such a stroke is unnecessary because the shorthand system is based on sounds, not standard spellings, and anyone familiar with the system having the knowledge of the content of what was said during the transcript phase, can easily determine where the words end. For the computer, to make such a determination, however, is more complex !

The problem of transcribing steno into written text raises difficulties that occur also in speech recognition, such as homonyms desambiguation and word boundaries determination. This concern was at the origin of the software called TASF (« Transcription Automatique de la Sténotypie en Français ») developed by the IBM France Scientific Center in Paris.

4. AUTOMATIC TRANSCRIPTION BY TASF

For English, the steno method is very precise, and the number of ambiguities is small. That explains why CAT have been developed for several years in English, with high performances.

For French, because of the confusions allowed by the method, and the inherent complexity of French vocabulary composed of many homophones, the average number of words that may start at each steno stroke is 10. In such conditions, a complex algorithm to choose among these possibilities is necessary.

For example, the list of possible words for the sequence of stenograms *"KAN NOU A FON"* is :

KAN	quant, quand, gants, gant, camps, camp, qu'en, qu'ans, qu'an.
NOU	nous, noues, nouent, noue, noue, n'houe, n'houes, n'ou, n'où, n'houx, nous, nouas.
A	as, a, à, avons.
FON	vont, font, fonds, fond.

TASF is a stenotypy to French transcription system based on a steno-French dictionary composed of 250,000 words and a language model for French. It was mainly designed for off-line transcription of reporting.

Each entry contains one spelling, the corresponding steno code, the possible parts-of-speech and the frequencies. The language model is based on the theory of Markov source already used in speech recognition systems [5], and allows to compute the probability of any sequence of words from the probability that a part-of-speech appears after two given parts-of-speech. This language model was built using a 1.2 million of words corpus.

From a sequence of stenograms, a left-to-right sub-optimal search algorithm finds the most probable sequence of words among all possible ones found in the dictionary. In this decoding process, a word is fixed according to a look-ahead of 4-6 words. This system is detailed in [2]. On

the average, at each syllable, 10 possible words could be found in the dictionary, that may start here and match some substring of the steno. To give a comparison of the efficiency of the model, a simple longest match method, without language model, gives 50% error rate on words [5].

In addition to that, TASF is composed of a simple user interface and a 200 words user dictionary (this limitation has been decided at this time because of DOS memory management).

5. FROM TASF TO MEDIATEX-TASF

TASF was intensively evaluated during 4 months. 250 hours of broadcast were used as a test-field by a qualified stenotypist using the version 1.0 of the system along with the original Grandjean method. First results were speaking by themselves :

- real error rate was more than 15%, which represents more than 100 invalid words per page,
- the user dictionary was too small (200 words maximum) to accommodate a single public affair show which uses many proper names,
- the system was providing too much word ambiguities leading to funny parallels ("Ministre des tas" vs "Ministre d'État").

5.1 Grandjean method modifications

Grandjean method was originally designed to accelerate the speed of keystroke hits, to keep up with fast, or even moderately paced, speakers. As long as a human being was reading the stenograms to convert them into a written text, there was no major problem. Disambiguation of similarly coded words was easily processed by human being supplying the correct spelling based on context. But, since now this task is devoted to be performed by a "stupid" computer, ambiguity has to be reduced.

Thus, we decided, along with Grandjean Institute, to modify the stenotyping training method. At the first time, it was not so simple to accept, but results are now so obvious that this new training method is to be widely used.

Five major modifications has reduced dramatically the ambiguities of the original method. Let's take as examples:

① Discrimination of voiced and unvoiced consonants by the mandatory use of "*" as an extra-stroke. For example : *KOU* = coup and *K*OU* = goût;

② Use of "*" for distinguishing sounds "s", "z" and "x"

③ Even at the end of a word, sound "che" is always coded SK.

5.2 Middle-dictionaries for words and expressions

First, since the translations will appear on-the-air instantly, without opportunity for proof-reading or editing, it is necessary that the real-time captioner be extremely accurate. This means not only avoiding typos but also writing homonyms in different ways. For example, in English the words "right," "write," and "rite" all sound alike. Since the steno system is based on the sounds of the words, and has nothing to do with spelling, the shorthand form for each of these three words could be the same. For real-time, each has to be written distinctly, since each word appears on the air immediately there is no opportunity to decide afterwards on the correct spelling.

Second, the computer can only translate the words that it knows. Real-time captioners must prepare for each broadcast by entering into the user's dictionary any words that are likely to come up in a real-time broadcast but which are not yet in the dictionary. For example, for a newscast, such items such as the names of all the newsmakers, the places in the news, any new medicines that may have been invented, and so on, must be put into the dictionary in advance of air time so that when they are spoken and transcribed during a broadcast they will translate properly.

Third, for a human being, it is well known that a foreign language is mastered when expressions are well known and properly used. Here, in a certain sense, it is the same: transcription errors are considerably reduced when the lexicon contains the longest possible character chain i.e full expressions such as "c'est à dire", "il y a ", "à bras le corps", "chanter en choeur", etc.

This task is achieved by specific lexicons called middle-dictionaries, containing names and expressions. At the present time, these lexicons are composed of more than 1,500 new words (proper names, locations and regionalisms) and 6,000 French expressions.

5.3 Post-processing rules

Statistical language modelling for unlimited vocabulary is known as being a very efficient method. Unfortunately, it is not error-free. For this purpose, 500 post-processing rules are then applied on the character string to solve some remaining lexical problems as well as verb conjugation and word agreement in gender and number. These rules are rewriting rules written in C language. Here is some examples:

- if a plural substantive is followed by "êtes", and if "êtes" is not followed by "vous", then replace "êtes" by "aident" *(êtes and aident are stenotyped the same way).*
- if a "être" auxiliary is followed by "à" and then by "les", replace both of them by "allé" agreed to the verb subject.

(ex: "ils sont à les vendre" -> "ils sont allés vendre").

6. TOWARDS A BETTER LANGUAGE MODEL

In order to improve transcription quality of spoken language, since the previous statistics for building the language model were derived from a large database originated from written text, we decided to re-train the language model using spoken language based contents. A new corpus of 687,000 words was collected during summer 1991. It was mainly composed of transcriptions from public affair TV shows, interviews and reports from the House of Commons (Hansard). The written form of this corpus was manually corrected before training.

Eventually, a new language model was computed. In addition to that, probabilities of the general model were corrected in terms of frequencies.

Since the corpus was specifically based on spoken-language, we expected to get a better accuracy of the grammatical model, since it was supposed to reflect the real "oral syntax". Unfortunately, results was not as good as expected. The major reason leads probably in the corpus size, which was too small for that purpose.

7. RESULTS

7.1 MEDIATEX-TASF structure

The system runs presently DOS on two PC 386 at 33 MHz, but will eventually run on a single 486 under OS2. The first PC is devoted to TASF task, the second runs the rules and formatting software. A serial link connects the two units.

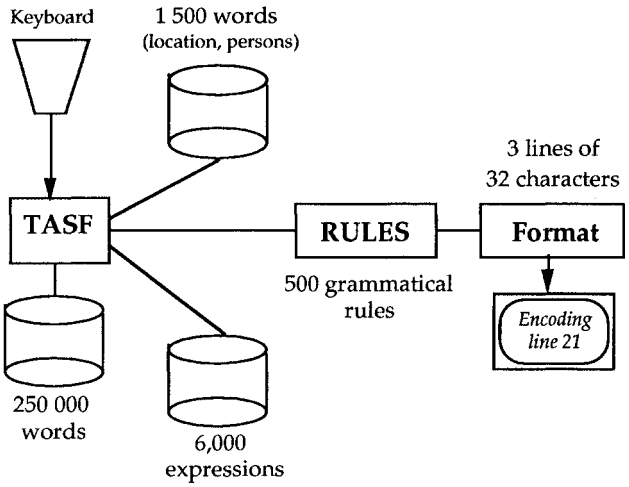

Fig. 2 - MEDIATEX-TASF Configuration

7.2 Assessment methodology

For measuring the error rate of MEDIATEX-TASF, one must have the original text as well as the results of the whole transcription process. An objective measurement tool was carried out specifically for this purpose. It was designed by a professional proof-reader who organized errors into 12 categories such as: errors in digits and numbers, in verb agreement, in typos, in homophony confusion, in punctuation, etc. Then, the different versions of TASF (from 1.0 to 1.5) were compared using this measuring tool. The results are the following:

Grandjean method modifications and the introduction of the two middle-dictionaries for words and expressions are contributing to the error rate decrease for 9%. In addition to that, 500 post-processing rules are contributing for 4% more. At last, the final error rate achieved after testing more than 1,000 hours of programs is less than 5%. In English, long experience centers (such as National Captioning Institute) are providing service with an error rate of 3% or better.

7.3 Global results

As we've seen, despite the incredible accomplishment of being able to translate in real-time spoken language by using MEDIATEX-TASF, the system is not a perfect one.

First of all, there is a delay between the time a word is uttered and the time the word appears on the screen. This delay is caused by the inability of a stenotypist to transcribe a word before it is heard and also by language modelling process that needs to have 3 words at least to decide the most probable word to appear in the string. For thi reason, real-time captions appear in a continuous "roll-up mode on 3 lines at the bottom of the screen.

Secondly, despite all the preparation, some erroi appear into the captions. Errors may be caused by th stenotypist simply hitting a wrong key, thus making typ But the result is here more important than on a typewrite because producing different shorthand codes from the on intended could result totally different word. Because c the algorithm used, the computer may misanalyze whic strokes go together, resulting in a word, or in a phrase, tha is phonetically similar to the intended one, but clearl wrong. And this could provide funny results... Fc example, the phrase « ... parti racheter un livre » could b confused with « ... partira jeter un livre ». The only way t prevent these errors is a painstaking task: after ever program, real-time captioners routinely have to g through what they have done, determine the origin c every errors, and take corrective action, such as adding new entry in the dictionaries of expressions or words, c modifying post-edition rules.

8. CONCLUSION

MEDIATEX-TASF was officially unveiled in Montrea on April 14 by M. Guy Gougeon, Vice-President of th CBC's French network. Then, it was presented to th international community of handicapped people on Apr 23 at the Independence '92 Conference in Vancouver. A last, MEDIATEX-TASF was used for captioning 12 hours of national meeting in Paris on June 19-20, during th "Assises Nationales des Étudiants Handicapés".

During next year, all the transcriptions will be kep from Radio-Canada broadcasting services. Since all th transcriptions will be corrected for improving the syster this will lead eventually to collect and record a very larg corpus of data, which could be one of the biggest databas available of French spoken language transcripts.

Acknowledgement

Parts of this text are excerpted from M. Okrand and from A.M. Derouault & B. Mérialdo previous papers.

References

[1] F. Néel, G. Adda, C. Chesnot, C. Fournier, C. Fluhı "Problèmes liés aux sous-titrages d'émission télévisées avec léger différé", *Colloqu francophone sur la technologie au service de personnes handicapées (HANDITEC)*, Dec. 1985.

[2] A.M. Derouault, B. Mérialdo, J.L. Stehle, "Un expérience de transcription automatiqu sténotypie-Français", *TSI*, Vol. 2, n. 5, Sept. 1983.

[3] A.M. Derouault, B. Mérialdo, "TASF: a Stenotypy-to French Transcription System", *ICASSP*, San Diego 84

[4] "Méthode de sténotypie", *Ed. Sténotypie Grandjean* 15 rue Soufflot, 75240 Paris Cedex 05 (FRANCE).

[5] A.M. Derouault, B. Mérialdo, "Language modeling a the syntactic level", *7th ICPR*, Montreal, 1984.

[6] M. Okrand, "Closed Captioning in Real Time" *SMPTE Journal*, pp. 427-432, June 1991.

* * *