Transcription manuelle *vs* assistée de la parole préparée et spontanée

Thierry Bazillon, Yannick Estève, Daniel Luzzati

LIUM (Laboratoire d'Informatique de l'Université du Maine)
Avenue Laënnec, 72085 Le Mans, France
prenom.nom@lium.univ-lemans.fr
http://www-lium.univ-lemans.fr

ABSTRACT

Our paper focuses on the gain which can be achieved on transcription of spontaneous and prepared speech, by using an ASR system. This experiment has shown interesting results, first about the duration of transcription task itself: even with the combination of prepared speech + ASR, an experimented annotator needs approximatively 8 hours to transcribe 2 hours of audio data. Then, using an ASR system is mostly time-saving, although this gain is much more significant on prepared speech: assisted transcriptions are up to four times faster than manual ones. This ratio falls to two with spontaneous speech, because of ASR limits for these data. Lastly, spelling correction is very time-consuming with prepared speech, because it contains many proper nouns that had to be checked; their frequency may be a way to detect spontaneous speech.

Keywords: spontaneous speech, prepared speech, transcription, time gain, ASR systems.

1. INTRODUCTION

Transcrire la parole est une tâche complexe : procéder de façon entièrement manuelle est très coûteux en temps, tandis que les systèmes de reconnaissance automatique ne sont pas encore assez performants si l'on souhaite des transcriptions très précises. La transcription assistée apparaît donc comme un bon compromis pour obtenir de tels résultats dans un temps raisonnable. Dans cette étude, notre but est donc de quantifier le gain qui peut être obtenu, pour la parole préparée et spontanée, entre des transcriptions manuelles et assistées. Ces dernières ont été réalisées à l'aide de LIUM RT, un système de reconnaissance automatique de la parole fondé sur le décodeur CMU Sphinx [4]. L'apprentissage de LIUM RT a été grande partie réalisé grâce à des articles issus du journal *Le Monde*, ce qui signifie que ce système n'est par essence pas optimisé pour reconnaître de la parole spontanée, telle qu'elle apparaît dans les débats ou les interviews. Ce travail est étroitement lié au projet EPAC (http ://epac.univ-lemans.fr), sélectionné par l'ANR dans le cadre de l'appel à projets 2006 du programme « Masse de Données - Connaissances Ambiantes »[1]. Outre le LIUM, qui en est le coordinateur, les laboratoires d'Avignon (LIA), de Tours (LI) et de Toulouse (IRIT) y sont également associés. Ayant débuté en mars 2007, EPAC a pour principal champ de recherche l'extraction et le traitement de la parole spontanée dans de grands corpus audio. Plus spécifiquement, les marqueurs acoustiques, l'identification nommée, la relation entre systèmes de reconnaissance automatique et parole spontanée, ou encore la transcription et l'annotation en sont quelques axes d'étude. L'un des objectifs d'EPAC est de fournir, d'ici mars 2009, les transcriptions annotées d'une centaine d'heures d'enregistrements radiophoniques issus de la campagne ESTER [5]. Le corpus ainsi constitué contiendra principalement de la parole spontanée, ce type de données faisant actuellement défaut à la communauté parole francophone.

Outre l'intérêt scientifique, c'est donc également en vue d'optimiser cette tâche de transcription que la présente étude a été menée.

[1] Ces travaux sont financés par l'ANR sous le contrat numéro ANR-06-MDCA-006.

2. PAROLE SPONTANEE *VS* PAROLE PREPAREE

La différence majeure entre ces deux types de paroles est le nombre de disfluences [1], en général bien plus présentes dans la parole spontanée. En conséquence, celle-ci est mal reconnue par les systèmes de reconnaissance automatique. Cependant, un précédent travail sur la parole spontanée et préparée [6] a montré que cette distinction pouvait être ambiguë : certains propos spontanés (par exemple lorsque les locuteurs sont des politiciens) ressemblent fortement à de la parole préparée ; à l'inverse, un discours préparé émaillé de plusieurs faux départs ou répétitions peut sembler spontané.

Ce travail se proposait de considérer la qualité d'élocution plutôt que l'opposition parole préparée / parole spontanée. Un corpus radiophonique d'environ onze heures a été choisi, comprenant des extraits de France Inter, France Info, Radio Classique, RFI et France Culture. Les fichiers ont été segmentés de façon automatique, et le texte était quant à lui supprimé. Deux annotateurs étaient chargés de noter chaque segment de parole suivant une échelle numérique allant de 1 à 9. 1 était la note attribuée à un segment sans aucune disfluence, avec une élocution parfaitement claire ; 9 indiquait un segment inaudible tant les hésitations, répétitions, faux départs... étaient

nombreux - ce cas extrême n'a jamais été rencontré au cours de l'expérience. Une note globale était ensuite attribuée à chaque tour de parole, pour éviter d'accorder la même importance à un segment très bref, donc potentiellement moins susceptible de comporter des disfluences, et à un segment long de plusieurs secondes.

Une partie de ces dix heures a été évaluée conjointement par les deux annotateurs, pour être sûrs qu'ils utilisaient bien les mêmes critères de notation. Ensuite le coefficient Kappa [3] a été calculé pour valider le processus. Un score de 0.852 a été obtenu, sachant que les scores dépassant 0.81 sont considérés comme étant excellents. Cela prouve que malgré la relative subjectivité du concept de « qualité d'élocution », les deux annotateurs étaient d'accord pour déterminer ce qui était de la parole de bonne qualité et ce qui n'en était pas.

Cependant, pour l'étude que nous nous proposons de mener ici, les fichiers de test ont été choisis selon la distinction conventionnelle spontané / préparé : ce qui est considéré comme préparé est du « broadcawst news », et par « spontané » nous entendons des débats ou des interviews. Par ailleurs, notre étude nécessitait des fichiers d'environ 10 minutes, et il est impossible de trouver dans un tel intervalle des segments ayant tous la même qualité d'élocution. Distinguer la parole spontanée – et tous les phénomènes que cela implique (faux départs, troncations, répétitions, parole superposée, morphèmes comme « euh », « ben »...) de la parole préparée est suffisant pour permettre à notre expérience d'obtenir des résultats et d'ouvrir des perspectives.

3. PROTOCOLE

Les données utilisées sont les suivantes : 24 segments d'environ 10 minutes, sélectionnés parmi les données non transcrites du corpus ESTER ; 12 sont considérés comme étant de la parole spontanée (débats ou interviews), et 12 comme de la parole préparée (informations). France Info n'a pas été pris en compte pour la parole spontanée, car le caractère même de cette station fait qu'elle ne contient que très rarement ce type de données. Par ailleurs, quand notre corpus comprenait des éléments non pertinents (musique...), ils n'ont pas été pris en considération. À l'aide de TRANSCRIBER [2], une transcription manuelle et une transcription assistée ont été effectuées sur chacun de ces fichiers par le même transcripteur (suffisamment longtemps après pour que la seconde transcription soit aussi peu influencée par la mémoire de la première que possible). Pour la transcription assistée, l'annotateur bénéficiait du texte brut généré par LIUM RT, ainsi que d'une segmentation automatique (donc susceptible d'être erronée) en tours de parole et locuteurs, sans identification nommée. Ainsi, il a semblé pertinent d'effectuer un chronométrage à la minute sur trois niveaux :

- la transcription et la segmentation en tours de parole

- l'assignation des locuteurs

- la vérification orthographique

4. PRINCIPAUX RESULTATS

Table 1 : *Durée totale de la transcription (durées respectives des corpus : 2H08 et 2H10)*

	Parole préparée	**Parole spontanée**
Transcription manuelle	17h36	19h33
Transcription assistée	8h31	15h44

La table 1 montre que la transcription assistée induit un important gain de temps, surtout pour la parole préparée. Pour ce type de données, le temps nécessaire à la transcription est approximativement deux fois moins important lorsque le transcripteur est assisté. Lorsqu'il s'agit de parole spontanée, ce bénéfice est bien moindre. À titre de comparaison, si l'on considère un segment de parole préparée de 10 minutes, le transcripteur aura besoin d'environ 40 minutes pour transcrire le texte, assigner les locuteurs et vérifier l'orthographe, s'il s'appuie sur un fichier de transcription généré automatiquement. Si l'on réalise les mêmes tâches sur le même fichier, mais de façon manuelle, environ 83 minutes seront nécessaires, soit un temps de travail plus que doublé.

La même expérience, mais cette fois avec un fichier de parole spontanée, montre qu'une transcription assistée demande 73 minutes de travail. À l'inverse, la transcription manuelle (90 minutes) n'est cette fois pas beaucoup plus coûteuse en temps que la transcription assistée. Ainsi, s'il est indéniable qu'une transcription assistée est synonyme de gain de temps, ce dernier est beaucoup plus important lorsqu'il s'agit de parole préparée.

Table 2 : *Transcription du texte et segmentation*

	Parole préparée	**Parole spontanée**
Transcription manuelle	13h36	16h15
Transcription assistée	5h06	12h41

C'est lors de la tâche de transcription du texte (Table 2) que le gain le plus intéressant a été obtenu : sur de la parole préparée, une transcription manuelle nécessite environ 2,67 fois plus de temps qu'une transcription assistée (5h06 *vs* 13h36). Pour être plus précis, le fichier

le plus significatif a obtenu un score de 3,75 : pour une durée effective de 00''08''55 (*hh/mm/ss*), la transcription assistée a ainsi demandé 00''14''49, et la transcription manuelle, 00''55''34. Ces chiffres sont très significatifs, notamment s'ils sont comparés à ceux obtenus avec la parole spontanée : pour une durée sensiblement équivalente, le rapport global chute à 1,28. Quant au fichier qui présente le gain le plus important, celui-ci n'est que de 1,95 : pour 00''11''18 de parole, la transcription assistée nécessite 00''47''03, et la transcription manuelle, 01''31''52.

Ces écarts mettent en exergue le fait que les système de reconnaissance automatique de la parole éprouvent des difficultés à traiter la parole spontanée, obligeant le transcripteur à effectuer par la suite beaucoup de corrections manuelles.

Table 3 : *Assignation des locuteurs*

	Parole préparée	**Parole spontanée**
Transcription manuelle	1h17	2h13
Transcription assistée	1h17	2h13

En ce qui concerne l'assignation des locuteurs (Table 3), les chiffres rigoureusement identiques que nous avons obtenus viennent confirmer une évidence : un système de reconnaissance automatique de la parole n'est d'aucune aide pour le traitement des locuteurs, dans la mesure où il ne les identifie pas nommément. Il est plutôt important de retenir que cette assignation demande presque deux fois plus de temps quand la parole est spontanée. Cela s'explique relativement facilement : la parole spontanée, avec ses nombreux tours de parole, contraint le transcripteur à leur assigner un locuteur, quand bien même il peut n'y en avoir que deux différents dans un fichier. À l'inverse un segment de parole préparée contient souvent de nombreux locuteurs (journalistes, reporters, interviewés, speakers...), mais beaucoup moins de tours de parole, dans la mesure où ceux-ci sont beaucoup plus longs. De plus, dans un segment spontané se trouve parfois de la parole superposée, et lorsque trois locuteurs ou plus sont susceptibles de prendre la parole, cela peut être long et difficile de déterminer qui parle réellement.

Table 4 : *Correction orthographique*

	Parole préparée	**Parole spontanée**
Transcription manuelle	2h43	1h05
Transcription assistée	2h08	0h51

Le minutage de la correction orthographique (Table 4) a permis d'observer un phénomène remarquable : si la différence spécifique entre transcription manuelle et assistée n'est certes pas très significative, celle entre parole préparée et spontanée l'est beaucoup plus. La raison en est fort simple : les segments de parole préparée contiennent essentiellement de l'information radiophonique ; or ce genre de données s'avère très riche en noms propres, assimilables à des entités nommées (reporters, interviewés, personnalités, villes...), dont les orthographes exactes ne peuvent être systématiquement connues de l'annotateur. Les rechercher peut donc être une tâche assez longue, notamment dans le cas de noms étrangers. Inversement, les fichiers de parole spontanée étant des interviews ou des débats, on y trouve très peu de noms propres car les thèmes abordés ne nécessitent en général qu'un faible recours aux entités nommées.

Table 5 : *Taux d'erreur mot (%)*

	Parole préparée	**Parole spontanée**
Transcription manuelle	16,95	35,21
Transcription assistée	15,83	34,33

Les dernières observations effectuées concernent le taux d'erreur mot (Table 5). Celui-ci a été mesuré à partir des sorties automatiques générées par le système LIUM RT, que nous avons comparées aux transcriptions manuelles puis assistées réalisées par l'annotateur. Les moyennes indiquées ci-dessus confirment ce que nous disions précédemment : le système de reconnaissance automatique du LIUM n'est pas aussi performant sur la parole spontanée que sur la parole préparée. Les différences observées entre les tâches manuelles et assistées peuvent être expliquées par le fait que le transcripteur n'a pas forcément transcrit le même texte à chaque fois : il est parfois difficile de percevoir clairement des phénomènes tels que les répétitions, les faux départs ou encore la parole superposée, et en conséquence leur

transcription ne sera pas toujours identique, même lorsqu'elle est réalisée deux fois par la même personne.

Des résultats plus détaillés montrent que le segment ayant obtenu le plus faible taux d'erreur mot est celui qui a nécessité le moins de temps pour être transcrit. De même, il représente le gain le plus important entre la tâche de transcription manuelle et assistée du texte (0h56 *vs* 0h15). Toutefois, le segment avec le taux d'erreur mot le plus fort (54,5%) n'a quant à lui pas demandé plus de temps qu'un autre pour être transcrit ; mais son principal locuteur est un photographe âgé s'exprimant de façon très sporadique, d'une voix presque chuchotée et sans articuler, ce qui explique en grande partie l'importance de ce pourcentage. Ainsi, dans notre corpus, ce fichier est l'un des seuls pour lequel la transcription assistée a demandé à peu près le même temps de travail que la transcription manuelle (0h58 *vs* 0h57), parce qu'effacer les formes erronées puis ré-écrire le texte exact s'avère être un très long travail quand le taux d'erreur mot est élevé.

Enfin, si l'on considère la tâche de transcription assistée, on s'aperçoit finalement que le taux d'erreur mot est en adéquation avec le rapport entre durée totale de la transcription et durée totale des fichiers (Table 2) : un transcripteur professionnel a approximativement besoin de deux fois plus de temps pour corriger un segment spontané qu'un segment préparé, et le taux d'erreur mot est approximativement deux fois plus important avec la parole spontanée qu'avec la parole préparée.

5. CONCLUSION ET PERSPECTIVES

Ce travail montre que la transcription, qu'elle soit réalisée de façon manuelle ou assistée, est un travail qui nécessite beaucoup de temps. Les meilleurs résultats que nous ayons obtenus sont ceux qui combinaient parole préparée et transcription assistée, mais même dans ce cas de figure, transcrire un corpus de 10 heures signifie approximativement 40 heures de travail. Quant à la transcription manuelle de 10 heures de parole spontanée, elle en demanderait le double. Notre étude illustre également le fait que, hormis quelques cas isolés, recourir à un système de reconnaissance automatique de la parole permet de gagner du temps, quand bien même ce gain peut varier considérablement d'un fichier à l'autre.

L'assignation des locuteurs peut sembler relativement coûteuse en temps sur la parole spontanée, mais il faut nuancer ce constat : en règle générale, transcription, assignation des locuteurs et correction orthographique sont effectuées conjointement par l'annotateur, et non de façon consécutive. En ce qui concerne les locuteurs, cela a son importance car le fait de les assigner après avoir transcrit le texte contraint le transcripteur à « relire » son travail depuis le début. Pour la parole préparée, cette relecture prend assez peu de temps puisque les tours de parole sont généralement longs et clairement distincts. A l'inverse, la parole spontanée contient fréquemment des tours de parole brefs, avec de nombreux changements de locuteurs ; ainsi, le transcripteur doit réécouter presque chaque segment séparément, ce qui demande beaucoup plus de temps.

La fréquence des entités nommées est une donnée intéressante dans la perspectives de futurs travaux : elle pourrait être un moyen ontologique de détecter la parole spontanée, ce qui s'avérerait très utile et permettrait un important gain de temps, notamment en ce qui concerne de grands corpus tels que ceux du projet EPAC.

BIBLIOGRAPHIE

[1] Adda-Decker, M. & alii (2004), « Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage », JEP 2004, Fès (Maroc), 19-22 avril 2004.

[2] Barras, C. & alii (1998), « Transcriber: a free tool for segmenting, labeling and transcribing speech », in *Proc. First Int. Conf. on language ressources and evaluation*, LREC 98, Grenade (Espagne), 28-30 mai 1998, pp. 1373-1376.

[3] Cohen J. (1960), « A coefficient of agreement for nominal scales », in *Educational and Psychological Measurement* 20, pp. 37-46.

[4] Deléglise, P. & alii (2005), « The LIUM Speech Transcription System: A CMU Sphinx III-Based System for French Broadcast News », Interspeech 2005, Lisbonne (Portugal).

[5] Galliano, E. & alii (2005), « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast New », in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Interspeech 2005, Lisbonne (Portugal).

[6] Jousse, V. & alii (2008), « Caractérisation et détection de parole spontanée dans de larges collections de documents audio », JEP 2008, Avignon (France).