

# Étude de l'informativité des transcriptions : une approche basée sur le résumé automatique

Carlos-Emiliano González-Gallardo, Malek Hajjem, Eric Sanjuan, Juan-Manuel Torres-Moreno

▶ **To cite this version:**

Carlos-Emiliano González-Gallardo, Malek Hajjem, Eric Sanjuan, Juan-Manuel Torres-Moreno. Étude de l'informativité des transcriptions : une approche basée sur le résumé automatique. Conférence en Recherche d'Information et Applications (CORIA), May 2018, Rennes, France. hal-01822585

**HAL Id: hal-01822585**
**https://hal.science/hal-01822585**

Submitted on 25 Jun 2018

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Étude de l'informativité des transcriptions : une approche basée sur le résumé automatique

**Carlos-Emiliano González-Gallardo**[*,**] **— Malek Hajjem**[*] **— Eric SanJuan**[*] **— Juan-Manuel Torres-Moreno**[*,**]

[*] *LIA, Université d'Avignon et des Pays de Vaucluse*
*339 chemin des Meinajariès, Agroparc BP 91228, F-84911 Avignon cedex 9, France*
*{carlos-emiliano.gonzalez-gallardo, malek.hajjem, eric.sanjuan, juan-manuel.torres}@univ-avignon.fr*
[**] *Département de Génie Informatique et Génie Logiciel, École Polytechnique de Montréal. 2900 Blvd. Edouard-Montpetit, H3T 1J4 Montréal (Québec) Canada*

RÉSUMÉ. *Dans cet article nous proposons une nouvelle approche d'évaluation de l'informativité des transcriptions issues de différents systèmes de Reconnaissance Automatiques de la Parole. Cette approche, fondée sur la notion d'informativité, s'inscrit particulièrement dans le cadre du Résumé Automatique de texte effectué sur ces transcriptions. Nous estimons, dans un premier temps, le contenu informatif des différentes transcriptions. Par la suite, nous explorons la capacité du Résumé automatique de texte pour surmonter la perte informative. Pour ce faire, nous utilisons un protocole d'évaluation automatique de résumés sans références basé sur le contenu informatif. Le but est de calculer les divergences entre les distributions de probabilité des différentes représentations textuelles obtenues: des transcriptions manuelles et automatiques et leurs résumés. À l'issue d'une série d'évaluations, cette analyse nous a permis de juger à la fois la qualité des transcriptions en terme d'informativité et d'évaluer la capacité du résumé automatique de texte à compenser les problèmes soulevés lors de la phase de transcription.*

ABSTRACT. *In this paper we propose a new approach to evaluate the informativeness of transcriptions coming from Automatic Speech Recognition systems. This approach, based in the notion of informativeness, is focused on the framework of Automatic Text Summarization performed over these transcriptions. At a first glance we estimate the informative content of the various automatic transcriptions, then we explore the capacity of Automatic Text Summarization to overcome the informative loss. To do this we use an automatic summary evaluation protocol without reference (based on the informative content), which computes the divergence between probability distributions of different textual representations: manual and automatic transcrip-*

*tions and their summaries. After a set of evaluations this analysis allowed us to judge both the quality of the transcriptions in terms of informativeness and to assess the ability of automatic text summarization to compensate the problems raised during the transcription phase.*

MOTS-CLÉS : *Mesures d'évaluation, Résumé automatique de textes, Transcription automatique, Métriques d'informativité*

KEYWORDS: *Evaluation measures, Text Automatic Summarization, Automatic Speech Recognition, Informativeness measures*

## 1. Introduction

Le résumé automatique de la parole constitue un cas particulier du résumé automatique de documents (RAD) (Torres-Moreno, 2014). Le flot d'entrée dans cette tâche est un flot de parole continue et la sortie sera sous la forme d'un résumé écrit ou parlé. L'une des méthodes la plus répandue de génération des résumés automatiques de la parole est de tirer parti des transcriptions automatiques du discours parlé (Ding *et al.*, 2012 ; Szaszák *et al.*, 2016 ; Taskiran *et al.*, 2001). Ce discours peut venir évidemment d'une vidéo. L'idée consiste donc à utiliser les algorithmes développés dans le cadre du résumé automatique de texte, afin de produire une version synthétique à partir de l'information contenue dans la parole présente dans les vidéos. Pour ce faire, une étape de transcription automatique du contenu parlé est nécessaire. Toutefois, dans ce cas, les traitements ne peuvent plus compter sur une transcription parfaite du texte à résumer, et doivent être capables de gérer les erreurs produites lors de l'étape de transcription. En effet, aux éventuelles erreurs du système de résumé automatique s'ajoutent les limites des systèmes de Reconnaissance Automatique de la Parole (RAP). Par conséquent, prévoir une stratégie pour estimer à quel point les méthodes issues du domaine textuel sont influencées par un contenu audio, devient indispensable. Ce travail de recherche s'inscrit dans ce cadre précis.

Nous visons l'évaluation de l'informativité d'un ensemble de résumés automatiques générés à partir des sorties des différents systèmes de transcription automatique. Les principales motivations de cette analyse s'inspirent essentiellement du besoin accru d'une représentation synthétique des informations parlées les plus importantes. En effet, de nos jours, les vidéos numériques représentent un facteur important pour véhiculer l'information. Avec l'avènement de plusieurs sites web d'hébergement vidéo tels que YouTube, Dailymotion, Hulu et autres ; l'utilisateur a la possibilité de regarder les émissions des chaînes TV à travers des podcasts. Cette rediffusion des émissions conduit à l'explosion du nombre de documents disponibles, ce qui rend nécessaire la gestion efficace de ce contenu vidéo dont le volume ne cesse pas d'augmenter.

Pour y arriver, une méthode de résumé par extraction est appliquée au contenu parlé, transcrit préalablement de façon automatique. Notre contribution concerne plus spécifiquement la phase d'évaluation de ces résumés obtenus. Nous cherchons, d'une part, à explorer la capacité d'un résumeur automatique à compenser la perte d'information issue des erreurs de transcription. D'autre part, nous cherchons à estimer l'influence du bruit généré par ces transcriptions sur les résumés automatiques. Ceci sera mesuré en termes du contenu informatif retenu à l'issue du processus d'extraction des passages les plus représentatifs.

La suite de l'article est organisée comme suit. Dans la section 2, nous présentons les travaux connexes. Dans la section 3, nous présentons les données utilisées dans le cadre de l'évaluation que nous proposons. La section 4 introduit une approche analytique qui vise à évaluer les résumés automatiques générés à base des transcriptions issues de différents systèmes RAP. Nous mettrons l'accent sur la mesure de l'informativité à travers des analyses poussées dont l'objectif est d'estimer la qualité des

transcriptions en terme du contenu informatif et de la capacité du RAD à compenser les erreurs de transcription, ainsi dans cette section nous discutons les résultats obtenus. Enfin, dans la section 5 nous présentons les conclusions de cet article.

## 2. État de l'art

### 2.1. *Reconnaissance automatique de la parole (RAP)*

La RAP est une démarche qui permet de passer d'un signal acoustique de parole à sa transcription dans une version écrite (Rabiner et Juang, 1993 ; Haton *et al.*, 2006). Elle consiste à employer des processus d'appariement afin de comparer une onde sonore à un ensemble d'échantillons, composés généralement de mots mais aussi des unités sonores encore plus élémentaires appelées phonèmes (Deng et Li, 2013).

On distingue deux approches pour la reconnaissances de ces formes élémentaires. La première est à base de règles utilisant des formes primitives préalablement déterminées par des experts (Lee et Kawahara, 2009). La réussite des expériences repose sur des conditions très strictes : vocabulaire réduit, mots isolés, peu de locuteurs, enregistrements dans des conditions de laboratoire (Lindgren, 1965). Ainsi, ces méthodes acoustiques seules sont insuffisantes et ont donné lieu à peu de réalisations concrètes car elles posent le problème de l'inférence des règles ainsi que celui de l'extraction des formes primitives (Haton *et al.*, 2006). En conséquence, des informations linguistiques commencent à être prises en compte dans les systèmes de reconnaissance pour ajouter des mots de contexte aux systèmes et ainsi améliorer la performance de la reconnaissance. Ceci a donné naissance à la deuxième approche qui consiste à extraire des vecteurs de paramètres caractéristiques à partir de ces formes afin d'utiliser une technique de classification permettant d'accorder une classe à une forme donnée (Baker, 1975 ; Jelinek *et al.*, 1975 ; Rabiner *et al.*, 1979). Dans ce cadre, les systèmes de reconnaissance de parole utilisent une approche statistique dites décodage statistique de la parole décrit dans (Jelinek, 1997).

Cette conversion parole-texte doit être indépendante de la taille du vocabulaire, de l'accent du locuteur, de son niveau de stress, etc. En effet, afin d'obtenir une transcription correcte, le système de transcription doit être capable de gérer les spécificités de la parole. Ainsi, il est évident que les performances des systèmes de RAP dépendent du type et de la qualité des données à transcrire (Galibert *et al.*, 2014). Généralement la performance d'un système de RAP est mesurée en termes de taux d'erreur de mots (*Word Error Rate*, WER). Le WER consiste à dénombrer les erreurs de transcription et à les normaliser par le nombre total de mots dans la référence pour fournir un pourcentage d'erreurs. Une telle métrique semble être efficace lorsque la transcription automatique est une fin en elle même. Cependant de nos jours un module de RAP est généralement combiné à plusieurs autres modules de Traitement Automatique du Langage naturel (TAL) afin de pouvoir résoudre des tâches encore plus complexes (Ben Jannet *et al.*, 2014), et le WER n'est plus adapté.

Le WER accorde le même poids à toutes les erreurs (erreurs d'insertion, d'omission ou de substitution). Cette stratégie d'évaluation entrave la prise en considération du cadre applicatif final. Dans ce contexte, un certain nombre de mesures alternatives ont été proposées dans la littérature. On cite la perte relative d'information (*Relative Information Loss*, RIL) (Miller, 1955), une mesure qui propose d'évaluer la perte d'information causée par les erreurs des systèmes de RAP. Cette métrique se base sur l'information mutuelle pour obtenir la force de la dépendance statistique entre le vocabulaire de la référence **X** et les mots de l'hypothèse **W**. Une approximation du RIL, c'est la perte d'information des mots (*Word Information Lost*, WIL) proposée par (Morris *et al.*, 2004). Cette métrique estime aussi la perte d'information due à des erreurs de transcription. Contrairement à la RIL, WIL prend en compte les mots corrects et les substitutions au moment de comparaison entre la référence et l'hypothèse. Une autre proposition de (McCowan *et al.*, 2004) consiste à reprendre les mesures de Recherche d'Information (RI) pour l'estimation de perte d'information causée par les dérives de la transcription. (Ben Jannet *et al.*, 2014) suggèrent une nouvelle méthodologie pour évaluer la qualité des transcriptions automatiques dans le contexte d'extraction d'entités nommées à partir de documents transcrits automatiquement. Ainsi ces différents travaux de l'état de l'art confirment la nécessité de revoir l'évaluation des transcriptions automatiques à base de WER et d'explorer d'autres types de métriques mieux adaptées pour mettre en évidence la qualité des transcriptions automatiques en termes d'informativité.

### 2.2. *Les mesures d'informativité et domaines d'application*

L'évaluation de l'informativité par identification de pépites informationnelles a été proposée dans un premiers temps dans (Nenkova et Passonneau, 2004). Cette méthode est basée sur la notion de *Summary Content Units* (SCU) ou pépites (*Nuggets*) définies manuellement par des annotateurs. Ces unités informationnelles, auxquelles les annotateurs accordent des poids, correspondent sensiblement au même contenu exprimé différemment. Un résumé automatique est dit informatif s'il est bien noté et qu'il contient des unités de fort poids.

Cette notion d'informativité a été étendue pour couvrir différentes briques technologiques issues de TAL. On la retrouve dans la tâche de traduction automatique en utilisant la méthode BLEU (*Bilingual Evaluation Understudy*) (Papineni *et al.*, 2002) et aussi sur l'évaluation des résumés automatiques en utilisant la méthode ROUGE (*Recall-Oriented Understudy for Gisting Evaluation*) (Lin, 2004). Le principe général de ROUGE est de calculer l'intersection des $n$-grammes communs entre le résumé automatique à évaluer et les résumés de référence disponibles. Pour une évaluation correcte avec ROUGE, les travaux de l'état de l'art ont montré qu'il est nécessaire de produire au moins cinq résumés de référence générés par différents annotateurs (Louis et Nenkova, 2009b). Ainsi, ROUGE s'avère inappropriée lorsque la génération des résumés de référence est trop coûteuse comme dans le cas de documents très longs ou

d'un très grand nombre de documents à résumer, ce qui représente le cas typique de la RI.

En effet dans le cadre de la RI cette notion de l'informativité à été reprise par la divergence de Kullback-Leibler (KL) (Kullback et Leibler, 1951). Cette mesure permet de comparer le contenu d'un résumé produit automatiquement à celui du document source. La comparaison est principalement basée sur l'étude des distributions de mots ou de ensembles de mots entre le résumé et les documents (Moriceau *et al.*, 2010). Une telle comparaison à base de distributions de probabilité est peu sensible à la présence (ou l'absence) de séquence de mots communs entre le résumé et la référence. Ceci la rend difficile à adapter pour les résumés guidés par une requête ou concernant un sujet particulier (Bellot *et al.*, 2015). Ce problème se concrétise dans le cadre de la contextualisation de tweets (SanJuan *et al.*, 2011), une tâche qui combine les notions de RI et de résumé automatique.

FRESA (*FRamework for Evaluating Summaries Automatically*)[1] est une méthode automatique inspirée des travaux de (Louis et Nenkova, 2009a) et de (Lin *et al.*, 2006) pour évaluer les résumés sans utiliser des références qui a été introduite par (Torres-Moreno *et al.*, 2010 ; Saggion *et al.*, 2010). La méthode intègre un prétraitement classique des documents (filtrage des mots non porteurs d'information, normalisation, etc.) avant de calculer la divergence des distributions de probabilités entre le document source et le résumé candidat. Ce prétraitement des documents permet de garder seulement les mots porteurs d'information et se focaliser sur l'informativité. Pour le calcul de la divergence $D(P||Q)$, FRESA a la possibilité de calculer la divergence de Jensen-Shannon ($\mathcal{JS}$) et également une modification de KL (Torres-Moreno, 2014) au moyen d'uni-grammes ($\text{FRESA}_1$), bi-grammes ($\text{FRESA}_2$), bi-grammes-SU4 ($\text{FRESA}_4$) et leur moyenne ($\text{FRESA}_M$).

FRESA a été utilisée pour évaluer la qualité des résumés de documents biomédicaux en langue catalane, où les résumés de référence des auteurs n'étaient pas disponibles (Vivaldi *et al.*, 2010). Egalement, FRESA a été employé lors de la campagne INEX 2010[2], volet question-réponse (*QA Track*, QA@INEX) (Torres-Moreno *et al.*, 2010). Cette dernière tâche combine les démarches de RI et du résumé automatique.

## 3. Corpus

Notre intérêt à l'analyse de l'informativité des transcriptions automatiques a été principalement motivé par l'explosion des données audiovisuels due à la rediffusion des émissions des chaînes TV à travers des podcasts d'actualités. Pour une meilleure visibilité du concept d'informativité, nous avons pris en compte le contexte multilingue à travers une collection de vidéos en français et en anglais.

1. FRESA est téléchargeable à l'adresse : `http://fresa.talne.eu`
2. `http://www.inex.otago.ac.nz/tracks/qa/qa.asp`

**Tableau 1.** *Thèmes des vidéos*

| Thème | Nombre de vidéos |
|---|---|
| Syrie | 6 |
| Territoires occupés | 4 |
| Donald Trump | 3 |
| Droits de l'homme | 3 |
| Terrorisme | 3 |
| Technologie | 1 |

Pour chaque langue nous avons sélectionné 10 vidéos à partir de l'hébergeur web YouTube pour les chaînes d'actualités : France24, RT, Euronews et BBC. Les documents audiovisuels de ce corpus ont été collectés par (Leszczuk *et al.*, 2017). Différentes thématiques ont été abordées dans le but de minimiser la prépondérance d'un thème sur la totalité des sujets abordés. Le tableau 1 illustre la distribution des thèmes du corpus.

En ce qui concerne la création du corpus textuel, une étape de transcription automatique du contenu parlé a été appliquée à travers de trois systèmes de RAP. En plus des sorties de ces trois systèmes, nous avons produit une transcription manuelle (dite transcription de référence, Réf-humaine) faite par différents experts maîtrisant la langue de la vidéo, afin d'avoir une ressource pour évaluer les performances des systèmes RAP. Pour une étude plus complète, nous avons décidé de tirer parti à la fois des systèmes RAP commerciaux et académiques.

La transcription automatique à base d'un système non commercial a été réalisée en utilisant KATS (*Kaldi-Based Transcription System*), système RAP qui a été introduit dans (Fohr *et al.*, 2017) et qui utilise des modèles acoustiques à base des réseaux de neurones profonds. En ce qui concerne les systèmes commerciaux, nous avons utilisé le système Google Cloud Speech API[3] (Google-ASR) ainsi que le système IBM speech-to text[4] (IBM-ASR) (Saon *et al.*, 2015). Ces deux derniers systèmes utilisent des modèles à base des réseaux de neurones combinés à d'autres statistiques.

La description du corpus de transcriptions (français et anglais) en termes de vocabulaire a été synthétisée au tableau 2. En effet, les vidéos les plus courtes impliquent un contenu textuel inférieur ou égale à 240 mots. En revanche les vidéos les plus longues excédent les 2 350 mots.

3. `https://cloud.google.com/speech`
4. `https://www.ibm.com/watson/services/speech-to-ext`

**Tableau 2.** *Statistiques des transcriptions*

| Langue | Système | Moyenne de mots | Ecart-type (mots) |
|---|---|---|---|
| Français | KATS | 950 | 704 |
| | Google-ASR | 847 | 644 |
| | IBM-ASR | 924 | 672 |
| | Réf-humaine | 958 | 702 |
| Anglais | KATS | 846 | 480 |
| | Google-ASR | 705 | 410 |
| | IBM-ASR | 870 | 510 |
| | Réf-humaine | 805 | 495 |

## 4. Mesure de l'informativité des transcriptions automatiques et l'impact du résumé automatique

Notre hypothèse est que le résumé automatique représente un moyen extrinsèque assez objectif pour évaluer la qualité des transcriptions venant d'un système RAP. On sait que l'informativité contenue dans un résumé vis-à-vis la source est un bon indicateur de la qualité d'un système de résumé automatique (Nenkova et Passonneau, 2004 ; Saggion *et al.*, 2010). Il est donc possible d'évaluer la qualité d'une transcription via la mesure d'informativité contenue dans le résumé correspondant.

Dans le cadre du résumé automatique, l'existence des phrases dans le texte source est essentielle pour repérer les phrases contenant les informations pertinentes. Les transcriptions issues des différents systèmes RAP ne contiennent pas des signes de ponctuation et représentent une séquence continue de mots. Ainsi, une étape de segmentation est nécessaire et représente un enjeu particulier en soi. Segmenter une transcription revient à établir des hypothèses des frontières de phrases en positionnant des signes de ponctuation dans la séquence initiale.

(González-Gallardo et Torres-Moreno, 2018) se sont intéressés en particulier à trouver les marqueurs de fin de phrase qui représentent les frontières des phrases en français. Nous nous sommes inspirés de ce travail pour la génération automatique des segments venant des transcriptions. Notre approche a été étendue dans le but de traiter la langue anglaise en plus du français. L'architecture à base de réseaux de neurones de convolution que nous avons appliqué sur les transcriptions en anglais et français est la même architecture qui (González-Gallardo et Torres-Moreno, 2018) ont signalé être la meilleure pour la segmentation des phrases en français.

Pour l'anglais et durant la phase d'apprentissage du réseau, nous avons utilisé un corpus de 426 millions de mots extraits de *English Gigaword Fifth Edition*[5]. L'évaluation du système a été appliquée sur un sous-ensemble de 106 millions de mots. Concernant le français, nous avons utilisé un corpus de 470 millions de mots extraits

5. `https://catalog.ldc.upenn.edu/LDC2011T07`

**Tableau 3.** *Performance des systèmes de segmentation des phrases*

| Langue | Précision | Rappel | F1 |
|---|---|---|---|
| Français | 0,845 | 0,754 | 0,795 |
| Anglais | 0,838 | 0,796 | 0,816 |

de *French Gigaword First Edition*[6] et un sous-corpus de 117 millions de mots pendant la phase d'évaluation. Le tableau 3 illustre la performance des deux systèmes en termes de précision, rappel et Fscore (F1) pour prédire les fins de phrases. Cette même stratégie de segmentation a été exécutée aussi bien pour les transcriptions automatiques que sur les transcriptions manuelles.

En ce qui concerne le résumé automatique, nous avons opté pour l'approche extractive, car en outre sa facilité d'implémentation, ses performances ont été bien établies dans les travaux de l'état de l'art (Torres-Moreno, 2014). Cette approche consiste à extraire, parmi les $P$ phrases qui constituent un document source, les $n$ phrases portant la plus grande quantité d'information. Ces phrases sont censées être les plus relevantes pour produire un résumé.

### 4.1. ARTEX

ARTEX (Autre Résumer de TEXtes) est un système de résumé automatique de documents par extraction de segments pertinents qui a été introduit par (Torres-Moreno, 2012). La première phase d'ARTEX consiste à faire un prétraitement du texte source. Il s'agit d'une étape très importante qui permet de normaliser les mots et de supprimer les mots outils (peu informatifs). Ceci avec le but de réduire la dimensionnalité de la représentation vectorielle et de pouvoir calculer l'informativité des phrases. Une fois le texte prétraité, le calcul de deux vecteurs type centroïde est réalisé : un vecteur lexical moyen $VL_j, j = 1...N$ (qui représente l'informativité moyenne du lexique de $N$ termes) et un vecteur thématique moyen $VT_i, i = 1...P$ (qui représente le thème central du document composé de $P$ phrases). Un produit scalaire de ces deux vecteurs (normalisé) est effectué pour chaque phrase $i$ du document du $i = 1...P$. Un poids pour chaque phrase sera donc obtenu moyennant une fonction à partir du produit scalaire. À l'issue de la pondération des phrases, la génération du résumé est simple : il consiste à concaténer les phrases ayant les scores les plus élevés dans l'ordre de leur occurrence dans le texte source. Une procédure de post-traitement (diminution de la redondance ou simplification) peut être appliquée à la fin du processus.

6. `https://catalog.ldc.upenn.edu/LDC2006T17`

Dans nos expériences, le ratio de compression[7] a été fixé à $\rho = 0.35$ pour obtenir les résumés automatiques. Le choix d'ARTEX pour la génération des résumé des transcriptions automatiques et manuelles est justifié par sa simplicité d'implémentation, sa rapidité d'exécution et ses résultats compétitifs (Morchid *et al.*, 2016). Nous notons aussi que ARTEX est assez indépendant des connaissances linguistiques, ce qui le rend particulièrement adapté à la proposition de résumer des transcriptions de documents audio en plusieurs langues[8].

### 4.2. *Évaluation de l'informativité*

La figure 1 illustre le protocole que nous avons suivi pour évaluer, d'abord l'informativité des transcriptions automatiques et ensuite l'impact du résumé automatique sur l'informativité.

Tout d'abord, les transcriptions manuelles et automatiques ont été effectuées sur les vidéos comme décrit dans la section 3. Ensuite, des résumés automatiques ont été générés en utilisant ARTEX. Enfin, pour mesurer l'informativité, nous avons calculé la divergence $\mathcal{KL}$ modifiée (Torres-Moreno, 2014) entre les distributions de probabilité des transcriptions manuelles et automatiques, ainsi que la divergence entre les distributions de probabilités des transcriptions manuelles et des résumés automatiques en utilisant la méthode FRESA.

Nous proposons trois scénarios d'évaluation basés sur les scores FRESA. Dans le premier scénario (Sc.1), nous comparons l'informativité entre la transcription humaine et les transcriptions automatiques produites par différents systèmes RAP. Puis, dans un deuxième scénario (Sc.2), l'informativité est mesurée entre la transcription humaine et les résumés produits à partir des transcriptions automatiques et manuelles. L'application de FRESA entre la référence humaine et son résumé établit une valeur maximale d'informativité attendue, qui est comparée à l'informativité des résumés venant des transcriptions automatiques. Enfin, pour un troisième scénario (Sc.3), nous comparons un ratio d'informativité de Sc.1 par rapport à Sc.2 afin d'évaluer la capacité du résumé à surmonter la perte d'informativité des systèmes RAP. Les tables 4 à 7 affichent les valeurs moyennes et l'écart-type des mesures FRESA. Nous avons considéré important de prendre en compte l'écart-type car cette valeur donne une idée générale de la façon dont l'informativité est influencée par les variations de longueur et de diversité des sujets traités.

7. Le ratio de compression représente le rapport entre la taille du résumé et la taille du document source (en nombre de phrases) : $\rho = \frac{|\text{résumé}|}{|\text{source}|}$.

8. Bien sûr, d'autres systèmes de résumé automatique statistiques ou non auraient pu été utilisés dans cette tâche. Il s'agit d'un module du type *plug-in*.

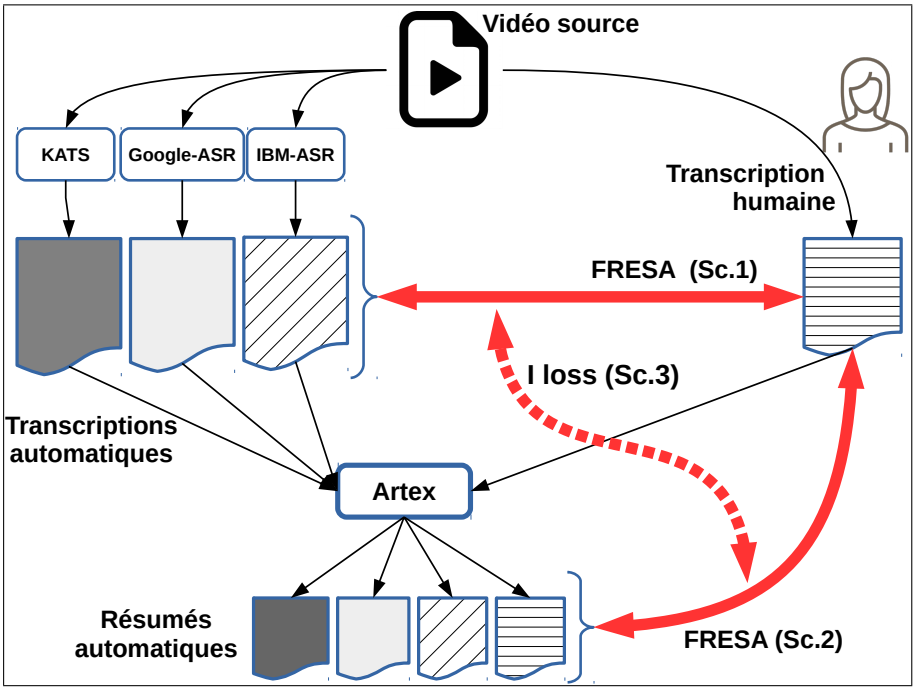



**Figure 1.** *Protocole d'évaluation utilisé lors de nos expériences. Scénarios d'évaluation : Sc.1 Scénario 1, Sc.2 Scénario 2 et Sc.3 Scénario 3.*

### 4.3. *Résultats*

– Transcription manuelle vs. Transcriptions automatiques (Sc.1)

Les scores FRESA pour le français et l'anglais concernant le premier scénario sont montrés dans les tableaux 4 et 5 respectivement. En français, on peut observer que le système KATS maintient un degré d'informativité supérieur sur toutes les mesures FRESA. L'informativité la plus basse est produite par le système IBM-ASR avec un score moyen ($\text{FRESA}_M$) de 0,539.

Un comportement différent est observé pour l'anglais. Google-ASR obtient la meilleure informativité sur presque tous les scores FRESA. Ceci est probablement du au fait que l'algorithme Google-ASR utilise plus de ressources pour l'anglais que pour le français. Un détail intéressant à signaler concerne le système IBM-ASR. Ce système produit le plus bas écart-type sur tous les scores FRESA, ce qui suggère qu'un léger mais stable degré d'informativité a été partagé sur les différents thèmes et les différentes longueurs.

**Tableau 4.** *Sc.1 - Transcription manuelle vs. Transcriptions automatiques (français)*

| SYSTEME | FRESA$_1$ | FRESA$_2$ | FRESA$_4$ | FRESA$_M$ |
|---|---|---|---|---|
| KATS | **0,835 ± 0,076** | **0,697± 0,118** | **0,683 ± 0,130** | **0,738± 0,106** |
| Google-ASR | 0,795 ± 0,132 | 0,664 ± 0,145 | 0,659 ±0,148 | 0,706± 0,140 |
| IBM-ASR | 0,662 ± 0,141 | 0,485 ± 0,134 | 0,471 ± 0,141 | 0,539 ±0,137 |

**Tableau 5.** *Sc.1 - Transcription manuelle vs. Transcriptions automatiques (anglais)*

| SYSTEME | FRESA$_1$ | FRESA$_2$ | FRESA$_4$ | FRESA$_M$ |
|---|---|---|---|---|
| KATS | **0,741 ± 0,061** | 0,584 ± 0,089 | 0,567 ± 0,094 | 0,631±0,080 |
| Google-ASR | 0,740 ± 0,093 | **0,605 ± 0,137** | **0,590±0,139** | **0,645 ±0,122** |
| IBM-ASR | 0,736 ± 0,058 | 0,578 ±0,076 | 0,566± 0,082 | 0,626±0,070 |

– Transcription manuelle vs. Résumés des transcriptions (Sc.2)

Les tableaux 6 et 7 montrent les résultats pour le deuxième scénario. La valeur maximale d'informativité attendue est établie par les scores basés sur les références humaines, et correspond au plus grand score FRESA qu'un résumé peut obtenir vis-à-vis de la référence manuelle. Plus le score FRESA du résumé automatique est proche de cette valeur, plus il sera informatif. Pour le français, Google-ASR possède le score moyen le plus proche $(0, 286)$ à la valeur d'informativité maximale attendue $(0, 395)$. Par contre, pour l'anglais, le score moyen d'informativité $0, 303$ a été obtenu par KATS, c'est-à-dire $0, 003$ plus haut qu'IBM-ASR.

Dans le tableau 2 on peut voir que Google-ASR a produit le plus bas nombre moyen de mots par transcription pour le français et l'anglais. Après une analyse manuelle de ces transcriptions, nous avons observé que lorsqu'une partie de l'audio dans la vidéo est difficile à comprendre, Google-ASR ne génère pas la transcription correspondante. Cela semble produire des effets opposés en fonction de la langue. Le processus de résumé a un impact positif pour le français, en éliminant de manière efficace les parties les moins informatives de la transcription. Ce comportement semble être à l'opposé en anglais, où l'informativité de Google-ASR passe de la première position de Sc.1 à la dernière de Sc.2. Nous pensons que l'excès moyen de mots transcrits que KATS et IBM-ASR génèrent par rapport à la transcription manuelle (tableau 2) influencent le score d'informativité de ces deux systèmes après le résumé.

**Tableau 6.** *Sc.2 - Transcription manuelle vs. Résumés (français)*

| SYSTEME | FRESA$_1$ | FRESA$_2$ | FRESA$_4$ | FRESA$_M$ |
|---|---|---|---|---|
| KATS | **0,385 ±0,080** | 0,238± 0,067 | 0,213 ± 0,069 | 0,279 ± 0,069 |
| Google-ASR | 0,377 ±0,098 | **0,249 ± 0,083** | **0,231 ±0,087** | **0,286 ±0,087** |
| IBM-ASR | 0,352 ± 0,076 | 0,200 ± 0,071 | 0,181 ±0,069 | 0,244±0,069 |
| Réf-humaine | 0,461 ±0,065 | 0,371 ±0,047 | 0,352 ±0,049 | 0,395 ±0,051 |

**Tableau 7.** *Sc.2 - Transcription manuelle vs. Résumés (anglais)*

| SYSTEME | FRESA$_1$ | FRESA$_2$ | FRESA$_4$ | FRESA$_M$ |
|---|---|---|---|---|
| KATS | 0,395 ± 0,078 | **0,266± 0,068** | **0,248 ±0,063** | **0,303 ±0,067** |
| Google-ASR | 0,342 ±0,085 | 0,222 ± 0,076 | 0,202 ± 0,069 | 0,256 ±0,075 |
| IBM-ASR | **0,396 ± 0,079** | 0,261 ±0,069 | 0,242 ±0,067 | 0,300±0,070 |
| Réf-humaine | 0,441 ± 0,042 | 0,347 ± 0,033 | 0,325 ±0,032 | 0,371 ±0,034 |

– Résumé automatique vs. Perte d'informativité (Sc.3)

Le scénario Sc.3 prend en compte les scores du scénario Sc.1 et du scénario Sc.2 afin de calculer la perte d'informativité produite par le résumé automatique. La perte d'informativité est exprimée comme un ratio entre le score FRESA de résumés venant de transcriptions automatiques et manuelles. Elle est donnée par l'équation [1].

$$I_{\text{loss}} = 100 \times \left(1 - \frac{\text{FRESA}_{x_{\text{système}}}}{\text{FRESA}_{x_{\text{Réf-humaine}}}}\right) \quad [1]$$

où $\text{FRESA}_{x_{\text{Réf-humaine}}}$ est égale à 1 pour le scénario Sc.1.

La figure 2 montre la perte d'informativité FRESA$_M$ en français et en anglais. On peut observer dans le cas du français, que les systèmes Google-ASR et IBM-ASR ont une perte plus petite après que le résumé automatique a été généré, avec une $I_{loss_{Sc.2}}$ qui diminue 1,8% et 7,94% respectivement. Par contre, le résumé automatique produit un impact négatif sur KATS, avec une augmentation de $I_{loss}$ de 3,27%. Un comportement distinct se présente en anglais. La figure 2 montre une augmentation de la performance en termes d'informativité. La perte de KATS tombe $18,61\%$, suivie par le système IBM-ASR $(18,15\%)$ et Google-ASR avec $4,41\%$.

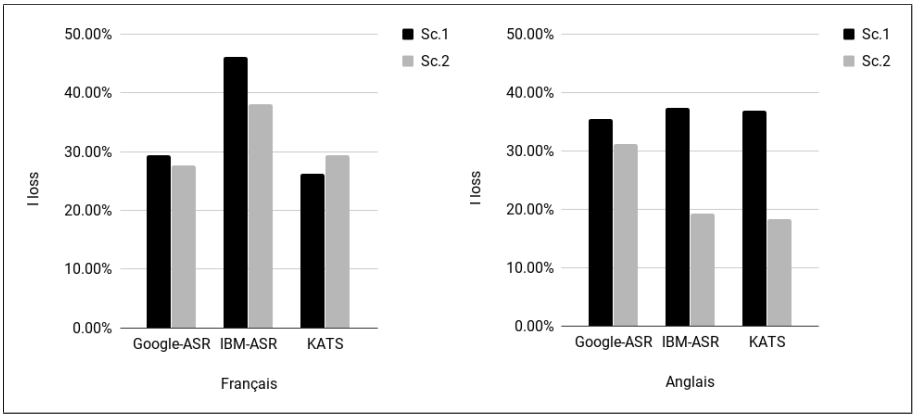



**Figure 2.** *Perte d'informativité pour l'anglais et français.*

## 5. Conclusion et perspectives

Dans cet article nous avons proposé une évaluation indirecte de la qualité des transcriptions au moyen d'une mesure de l'informativité dans le résumé automatique. Elle a été inspirée des travaux sur la recherche d'information ainsi comme du résumé automatique. Ce type de méthode d'évaluation produit une meilleure compréhension de la qualité des transcriptions automatiques en comparaison avec les mesures standards comme WER. À notre connaissance, il s'agit du premier travail qui introduit une telle mesure d'informativité dans l'évaluation de la qualité des transcriptions et de leur utilisation dans une chaîne de traitement automatique de langues.

Nous avons montré qu'en général, le résumé automatique peut augmenter le contenu informatif venant des transcriptions automatiques. Les résultats sont exploitables pour guider l'évaluation automatique (ou semi-automatique) des systèmes RAP couplés aux systèmes TAL. La démarche reste indépendante du domaine des documents et assez indépendante de la langue, bien que nos tests aient seulement été réalisés pour le moment en français et en anglais. Des tests en langue arabe sont actuellement en cours. En outre, nous allons augmenter le nombre de transcriptions manuelles afin de réduire le biais produit par leur nombre limité et également tester d'autres approches de résumé automatique.

Remerciements

Nous remercions le soutien financier du programme européen Chist-Era à travers du projet *Access Multilingual Information opinionS (AMIS)*, ANR-France / Europe.

## 6. Bibliographie

Baker J., « The DRAGON system–An overview », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, n° 1, p. 24-29, Feb, 1975.

Bellot P., Moriceau V., Mothe J., Éric SanJuan, Tannier X., « Mesures d'informativité et de lisibilité pour un cadre d'évaluation de la contextualisation de tweets », *Document Numérique*, vol. 18, n° 1, p. 55-73, 2015.

Ben Jannet M. A., Adda-Decker M., Galibert O., Kahn J., Rosset S., « How to assess the quality of automatic transcriptions for the extraction of named entities ? », *XXXe Journées d'Études sur la Parole (JEP'14)*, Le Mans, France, p. 430-437, June, 2014.

Deng L., Li X., « Machine learning paradigms for speech recognition : An overview », *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, n° 5, p. 1060-1089, 2013.

Ding D., Metze F., Rawat S., Schulam P. F., Burger S., Younessian E., Bao L., Christel M. G., Hauptmann A., « Beyond Audio and Video Retrieval : Towards Multimedia Summarization », *2Nd ACM*, ICMR '12, ACM, New York, NY, USA, p. 2 :1-2 :8, 2012.

Fohr D., Mella O., Illina I., « New Paradigm in Speech Recognition : Deep Neural Networks », *IEEE International Conference on Information Systems and Economic Intelligence*, 2017.

Galibert O., Leixa J., Gilles A., Choukri K., Gravier G., « The ETAPE Speech Processing Evaluation », *Language Resources and Evaluation*, Reykyavik, Iceland, May, 2014.

González-Gallardo C.-E., Torres-Moreno J.-M., « Sentence Boundary Detection for French with Subword-Level Information Vectors and Convolutional Neural Networks », *ArXiv*, February, 2018.

Haton J.-P., Cerisara C., Fohr D., Laprie Y., Smaïli K., *Reconnaissance Automatique de la Parole Du signal à son interprétation*, UniverSciences (Paris), Dunod, 2006.

Jelinek F., *Statistical Methods for Speech Recognition*, MIT Press, Cambridge, MA, USA, 1997.

Jelinek F., Bahl L., Mercer R., « Design of a linguistic statistical decoder for the recognition of continuous speech », *IEEE Transactions on Information Theory*, vol. 21, n$^{o}$ 3, p. 250-256, May, 1975.

Kullback S., Leibler R. A., « On information and sufficiency », *The annals of mathematical statistics*, vol. 22, n$^{o}$ 1, p. 79-86, 1951.

Lee A., Kawahara T., « Recent development of open-source speech recognition engine julius », *APSIPA ASC and IPA, 2009 Annual Summit and Conference*, Asia-Pacific Signal and Information Processing Association, p. 131-137, 2009.

Leszczuk M., Grega M., Koźbiał A., Gliwski J., Wasieczko K., Smaïli K., « Video Summarization Framework for Newscasts and Reports–Work in Progress », *International Conference on Multimedia Communications, Services and Security*, Springer, p. 86-97, 2017.

Lin C.-Y., « ROUGE : A Package for Automatic Evaluation of summaries », *Proc. ACL workshop on Text Summarization Branches Out*, p. 10, 2004.

Lin C.-Y., Cao G., Gao J., Nie J.-Y., « An information-theoretic approach to automatic evaluation of summaries », *Conference on Human Language Technology Conference of the North American Chapter*, ACL, Morristown, NJ, Etats-Unis, p. 463-470, 2006.

Lindgren N., « Machine recognition of human language Part III - Cursive script recognition », *IEEE Spectrum*, vol. 2, n$^{o}$ 5, p. 104-116, May, 1965.

Louis A., Nenkova A., « Automatically Evaluating Content Selection in Summarization without Human Models », *Conference on Empirical Methods in Natural Language Processing*, ACL, Singapour, p. 306-314, 6-7 août, 2009a.

Louis A., Nenkova A., « Performance Confidence Estimation for Automatic Summarization », *12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, ACL, Stroudsburg, PA, USA, p. 541-548, 2009b.

McCowan I. A., Moore D., Dines J., Gatica-Perez D., Flynn M., Wellner P., Bourlard H., On the Use of Information Retrieval Measures for Speech Recognition Evaluation, Idiap-RR n$^{o}$ Idiap-RR-73-2004, IDIAP, Martigny, Switzerland, 0, 2004.

Miller G. A., « Note on the bias of information estimates », *Information Theory in Psychology : Problems and Methods*, p. 95-100, 1955.

Morchid M., Torres-Moreno J.-M., Dufour R., Rodríguez J. R., Linarès G., « Automatic Text Summarization Approaches to Speed up Topic Model Learning Process », *Int. J. Comput. Linguistics Appl.*, vol. 7, n$^{o}$ 2, p. 87-109, 2016.

Moriceau V., SanJuan E., Tannier X., Bellot P., « Overview of the 2009 QA Track : Towards a Common Task for QA, Focused IR and Automatic Summarization Systems », *in* S. Geva, J. Kamps, A. Trotman (eds), *Focused Retrieval and Evaluation*, Berlin, p. 355-365, 2010.

Morris A. C., Maier V., Green P. D., « From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. », *INTERSPEECH*, ISCA, 2004.

Nenkova A., Passonneau R. J., « Evaluating Content Selection in Summarization : The Pyramid Method. », *HLT-NAACL*, p. 145-152, 2004.

Papineni K., Roukos S., Ward T., Zhu W.-J., « BLEU : a method for automatic evaluation of machine translation », *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, p. 311-318, 2002.

Rabiner L., Levinson S., Rosenberg A., Wilpon J., « Speaker-independent recognition of isolated words using clustering techniques », *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, n° 4, p. 336-349, Aug, 1979.

Rabiner L. R., Juang B.-H., *Fundamentals of speech recognition*, vol. 14, PTR Prentice Hall Englewood Cliffs, 1993.

Saggion H., Torres-Moreno J.-M., da Cunha I., SanJuan E., « Multilingual summarization evaluation without human models », *Proceedings of the 23rd International Conference on Computational Linguistics : Posters (COLING'10)*, ACL, Beijing, Chine, p. 1059-1067, 2010.

SanJuan E., Bellot P., Moriceau V., Tannier X., « Overview of the INEX 2010 Question Answering Track (QA@INEX) », *in* S. Geva, J. Kamps, R. Schenkel, A. Trotman (eds), *Comparative Evaluation of Focused Retrieval*, Springer, Berlin, Heidelberg, p. 269-281, 2011.

Saon G., Kuo H.-K. J., Rennie S., Picheny M., « The IBM 2015 English conversational telephone speech recognition system », *arXiv preprint arXiv :1505.05899*, 2015.

Szaszák G., Tündik M. Á., Beke A., « Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer. », *KDIR*, p. 221-227, 2016.

Taskiran C. M., Amir A., Ponceleon D. B., Delp E. J., « Automated video summarization using speech transcripts », *Storage and Retrieval for Media Databases 2002*, vol. 4676, International Society for Optics and Photonics, p. 371-383, 2001.

Torres-Moreno J.-M., « Artex is another text summarizer », *arXiv :1210.3312*, 2012.

Torres-Moreno J.-M., *Automatic Text Summarization*, Wiley and Sons, London, UK, 2014.

Torres-Moreno J.-M., Saggion H., da Cunha I., SanJuan E., « Summary Evaluation With and Without References », *Polibits : Research journal on Computer science and computer engineering with applications*, vol. 42, p. 13-19, 2010.

Vivaldi J., da Cunha I., Torres-Moreno J.-M., Velázquez-Morales P., « Generació automàtica de resums de textos especialitzats : experimentacions en llengua catalana », *Terminàlia*, vol. 1, p. 26-32, 2010.