# Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques

## Plan d'évaluation

## ESTER 2

## Phase 1

http://www.afcp-parole.org/ester/index.html

Version 0.3 du 17 janvier 2008

# Table des matières

# 1 Préambule

Ce document préliminaire décrit le **plan d'évaluation de la phase 1 de la deuxième campagne d'Évaluation des Systèmes de Transcription enrichie d'Émissions Radiophoniques** (ESTER 2). Il a pour objectif de définir les tâches traitées, les métriques d'évaluation utilisées et les données disponibles.

Cette nouvelle campagne d'évaluation s'inscrit dans la continuité des évaluations ESTER 1, menées entre 2003 et 2005. La campagne ESTER 2 reprend donc en grande partie les tâches des campagnes précédentes tout en proposant de nouvelles tâches exploratoires. Cette nouvelle campagne vise également à étendre progressivement les résultats des précédentes campagnes à d'autres types de données, notamment la parole accentuée et la parole spontanée. Comme l'édition précédente, la campagne ESTER 2 est prévue en deux phases. Le présent document décrit les règles d'évaluation pour la phase 1.

La campagne ESTER 2 vise plusieurs objectifs complémentaires. Le premier objectif est bien évidemment la mesure objective des performances et des progrès effectués depuis la campagne ESTER 1 pour les différentes composantes d'un système d'analyse et d'indexation de documents audio contenant de la parole. Un deuxième objectif est de produire des ressources (règles de transcription et d'évaluation et corpus annoté de taille conséquente) essentielles pour développer ou approfondir les recherches en transcription du français et d'en offrir l'accès à la communauté scientifique et industrielle la plus large possible dans le domaine. Plus généralement, la campagne éspère promouvoir une dynamique d'évaluation et d'échange au sein de la communauté de recherche et dégager de nouveaux axes de recherche à travers l'introduction de tâches exploratoires, notamment dans le domaine de l'extraction d'information et de l'enrichissement de la transcription.

# 2 Tâches de l'évaluation

La campagne ESTER 2 Phase 1 reprend en grande partie les tâches existantes lors de la précédente campagne d'évaluation. Ces tâches sont organisées autour de trois thèmes, à savoir, la segmentation (S), la transcription orthographique (T), et l'extraction d'information (E). Les tâches envisagées pour chacun des thèmes de la campagne sont résumées dans le tableau suivant et détaillées dans la suite de ce document.

| thème | tâche | description |
|---|---|---|
| S | SES | suivi d'événements sonores |
| S | SRL | segmentation en locuteurs |
| S | SVL | suivi de locuteurs |
| T | TRS | transcription orthographique |
| T | TTR | transcription temps réel |
| T | TDC | transcription avec données contemporaines |
| E | SP | structuration en unités syntaxiques (reponctuation) |
| E | EN | détection d'entités nommées |

Les tâches TDC et SP sont proposées à titre exploratoire. Par conséquent, la définition de ces tâches ainsi que les règles d'évaluation afférentes sont susceptibles d'être modifiées en concertation avec les participants, notamment à l'issue de la première phase de test. De plus, pour les tâches exploratoires, aucun classement des systèmes ne sera publié en dehors du cadre des participants.

## 2.1 Segmentation

Les tâches de segmentation visent à produire une structure du document selon certains critères. Elles sont basées sur la détection, le suivi et le regroupement d'événements sonores connus a priori ou pas. Trois tâches sont considérées dans cette catégorie :

- SES – suivi d'événements sonores
- SRL – détection des tours de parole et regroupement en locuteur
- SVL – suivi de locuteur

### 2.1.1 Suivi d'événements sonores (SES)

La tâche de suivi d'événements sonores consiste à détecter les événements suivants dans les documents sonores : musique (SES-M), parole (SES-P) et parole superposée (SES-2). Pour un événement donné, la réponse du système est l'ensemble des plages contenant l'événement considéré dans les documents de test. Les résultats seront retournés en un seul fichier au format ETF décrit en annexe de ce document.

Pour chaque événement, les performances seront mesurées en taux d'erreur sur la durée d'un document, comme défini dans la section 2.1.4. Cependant, à l'inverse des évaluations précédentes, les systèmes seront évalués séparément pour chacun des événements et un classement séparé sera etabli[1].

### 2.1.2 Segmentation et regroupement de locuteurs (SRL)

La tâche de segmentation et regroupement de locuteurs a pour but d'évaluer les algorithmes permettant de découper le flux audio en tours de parole tout en regroupant les plages associées à un même locuteur. L'identification des locuteurs concernés n'est pas requise. Le système retourne une segmentation du document spécifiant les plages de silence et, pour les zones contenant de la parole, une segmentation en locuteur où un identifiant arbitraire de locuteur est affecté à chaque segment. Pour chaque document du corpus de test, la segmentation et le regroupement devront être établis sur la seule base de ce document. Les résultats seront retournés en un seul fichier au format MDTM décrit en annexe de ce document.

L'utilisation de connaissances a priori, comme l'utilisation de modèles de locuteurs connus, est autorisée. Cependant, les participants utilisant des connais-

[1] À des fins de diagnostique, on envisagera cependant une métrique combinant les taux d'erreurs sur la musique et la parole, par exemple basée sur le taux d'erreur moyen.

sances a priori sont invités à soumettre un contraste où aucune connaissance a priori n'est utilisée.

### 2.1.3 Suivi de locuteurs (SVL)

Cette tâche vise à évaluer les systèmes de suivi de locuteur qui permettent l'enrichissement de la transcription ou de la description du document. L'objectif est de détecter les plages correspond à un locuteur donné, connu à l'avance.

Une liste des locuteurs cibles sera fournie avec les données d'apprentissage et de développement. Ces locuteurs sont nommément identifiés et possèdent un minimum de deux minutes de parole dans les données d'apprentissage. Les données d'apprentissage pour un locuteur donné consiste en l'ensemble des données prononcées par un locuteur cible dans le corpus d'apprentissage.

Il est à noter que la proportion de parole provenant de l'ensemble des locuteurs cibles ne sera pas nécessairement la même pour le corpus de développement et pour le corpus de test. Dans la mesure du possible, nous nous efforcerons de maintenir un écart faible entre ces deux proportions. Aucune garantie n'est donnée au niveau d'un locuteur cible particulier sur l'équilibre des proportions entre données de développement et données de test.

Les résultats seront fournis sous la forme d'un unique fichier au format ETF comme décrit en annexe de ce document. Les performances seront évaluées en terme de F-mesure moyennée sur l'ensemble des locuteurs cibles[2].

### 2.1.4 Mesure des performances

**Suivi d'événements.** Pour les tâches dont le but est la détection des plages contenant un événement sonore donné (SES et SVL), les performances seront mesurées par le taux d'erreur ramené à la durée du corpus, défini comme suit.

$$e_i = \frac{\text{omission}(i, \mathcal{C}) + \text{insertion}(i, \mathcal{C})}{T_{\mathcal{C}}} \tag{1}$$

avec

| | |
|---|---|
| $\mathcal{C}$ | ensemble des documents du corpus, |
| $T_{\mathcal{C}}$ | durée totale du corpus $\mathcal{C}$, |
| $\text{omission}(c_i, \mathcal{C})$ | temps où l'événement $i$ n'a pas été détecté dans $\mathcal{C}$, |
| $\text{insertion}(c_i, \mathcal{C})$ | temps où l'événement $i$ a été détecté à tort dans $\mathcal{C}$. |

Les temps seront mesurés en secondes, avec une tolérance de 0.25 secondes aux frontières des segments de la transcription de référence. Lorsque plusieurs événements sont considérés, les performances seront moyennées sur l'ensemble

[2]Il est à noter que cette métrique est différente de celle utilisée lors des précédentes évaluations, bien qu'également basée sur la F-mesure. La différence réside dans le fait qu'il s'agit ici de la moyenne des F-mesures par locuteur plutôt que de la F-mesure calculée sur l'ensemble des locuteurs cibles, cette dernière n'ayant pas d'interprétation intuitive.

des événements, soit un taux d'erreur moyen défini comme

$$e = \frac{\sum_{i \in \mathcal{E}} e_i}{\mathrm{card}(\mathcal{E})} \ , \tag{2}$$

où $\mathcal{E}$ désigne l'ensemble des événements considérés et $\mathrm{card}(\mathcal{E})$ le cardinal de cet ensemble[3]. Par exemple, dans la tâche de suivi de locuteur (SVL), $\mathcal{E}$ représente l'ensemble des locuteurs cibles et $e$ est donc le taux d'erreur moyen par locuteur cible.

**Segmentation en locuteur.** Pour la tâche de segmentation et regroupement de locuteurs (SRL), la métrique utilisée est le taux d'erreur défini comme la somme des taux de parole non détectée, de fausse détection de parole et de mauvaise détection. Le taux de parole non détectée correspond aux portions de parole détectée comme silence. Inversement, le taux de fausse détection de parole correspond aux portions de silence pour lesquelles un locuteur a été détecté. Le taux de mauvaise détection correspond aux erreurs sur les identités (arbitraires) des locuteurs. Une correspondance entre noms de locuteurs et noms arbitraires fournies par le système est établie par appariement. Ces taux seront calculés globalement sur l'ensemble des documents, sur la base du temps (en seconde) avec une tolérance de 0.25 secondes aux frontières des segments de référence.

**Outils de mesure des performances.** Pour l'ensemble des tâches de segmentation, les segmentations de références seront établies à partir des transcriptions manuelles pour les corpus de développement et de test. Les zones non transcrites dans la référence, indiquées dans un fichier au format UEM, ne seront pas prises en compte dans la mesure des performances. Par ailleurs, les conventions d'annotation stipulent que les pauses d'une durée inférieure à 0.5 secondes ne sont pas annotées comme silence dans les transcriptions de référence et ne seront donc pas considérées comme telles pour l'évaluation des tâches du thème S.

Une nouvelle version des progammes de mesure des performances et d'extraction des fichiers de référence[4] à partir des transcription au format `Transcriber` sera distribuée le plus rapidement possible après la distribution des données de développement.

[3] Par rapport aux métriques précédemment utilisées, le taux d'erreur tel que défini ici est relatif à la durée d'un document (ce qui revient donc à mesurer la quantité d'erreurs faites par le système par unité de temps) et non à la durée du temps de cible comme cela était le cas avec le rappel et la précision. Le recours à la moyenne sur les événements permet un lissage des résultats offrant ainsi les avantages d'une métrique qui mesure le comportement moyen du système.

[4] Pour la mesure de performances, seul le programme `trackeval` est concerné, le programme `SpkrSegEval-v23.pl` restant la référence pour la tâche SRL. Pour les fichiers de références aux formats ETF, UEM et MDTM, les outils `trs2etf`, `trs2mdtm` et `trs2uem` seront mis à jour.

## 2.2 Transcription

Ce thème vise à évaluer la transcription orthographique en sortie des systèmes de reconnaissance automatique de la parole, en terme de taux d'erreur de mots. Deux catégories de systèmes sont envisagées selon que le temps de calcul alloué au système est contraint (TTR) ou libre (TRS). Ces deux tâches principales, similaires à celles d'Ester 1, interdisent l'utilisation de données contemporaines aux données de test. En complément, la campagne Ester 2 propose une nouvelle tâche exploratoire qui autorise l'utilisation de données contemporaines aux données à transcrire afin de favoriser le développement de systèmes versatiles et adaptatifs (TDC).

Pour l'ensemble des tâches de transcription, est autorisée sans restriction l'utilisation de toutes les ressources acoustiques et linguistiques antérieures au 01/01/2008. Pour la tâche de transcription avec données contemporaines (TDC), des données linguistiques postérieures à la date du 01/01/2008 peuvent être utilisées pour l'adaptation non supervisée du système. La distribution d'un corpus de ressources linguistiques contemporaines en accompagnement des données de test est à l'étude.

Pour chacune des tâches, les résultats seront fournis sous la forme d'un fichier unique au format CTM décrit dans l'annexe B.

### 2.2.1 Transcription orthographique (TRS)

Cette tâche consiste à produire une transcription orthographique à partir du signal de parole, sans contrainte de temps de traitement autre que le délai global de la période de test.

La normalisation (adaptation au locuteur, normalisation de scores, etc) ne peut se faire que sur la base d'un document. Un contraste utilisant une normalisation sur plusieurs documents est toujours possible mais ne pourra être présenté comme système principal.

### 2.2.2 Transcription temps réel (TTR)

Cette tâche, similaire à la tâche TRS, comporte une contrainte additionnelle sur le temps de calcul pour le système complet qui ne doit pas être supérieur à une fois le temps réel sur un mono-processeur standard.

Le temps de calcul comprend l'ensemble des opérations (segmentation, E/S, décodage, etc.) et se mesure entre l'instant où le traitement est lancé et l'instant où il finit (commande `date` sous Unix). C'est le temps perçu par un utilisateur du système, et non le seul temps CPU (tel que le mesurerait la commande `time` sous Unix). Le traitement doit être lancé à froid (pas de lancement partiel pour mettre des informations en cache). Le ratio temps réel est le résultat de la division du temps de traitement par la durée du fichier, noté $X$xTR.

Se qualifie pour la tâche TTR un système capable de traiter le corpus de développement en moins de 1xTR en moyenne sur l'ensemble des fichiers. Le temps de traitement effectif sur les données de test pourra éventuellement dépasser 1xTR. On rapportera les temps de traitement non seulement pour le

test mais aussi pour le corpus de développement. On publiera les caractéristiques de la machine (Marque/type, CPU, bus, mémoire vive, disque dur). Les paramètres du système concernant le temps réel doivent être réglés de façon unique sur l'ensemble des fichiers et non fichier par fichier, que ce soit sur le corpus de développement (qualification à la tâche temps réel) ou sur celui de test.

### 2.2.3 Transcription avec données contemporaines (TDC)

La campagne ESTER 2 inaugure une nouvelle tâche, qualifiée d'exploratoire, concernant l'utilisation de ressources linguistiques contemporaines aux données à transcrire. Cette tâche vise à *évaluer les capacités d'adaptation non supervisée* d'un système à un nouveau type de données. Les systèmes pour la tâche TDC pourront utiliser des données contemporaines ou postérieures aux données à transcrire, de manière non supervisée. La contrainte de non supervision interdit en particulier le développement d'un système dont le modèle de langage initial (*i.e.*, avant utilisation des données contemporaines pour adaptation) inclurait des données contemporaines ou postérieures aux données de tests. En phase de test, aucun des paramètres du système ne devra être réglé manuellement sur la base des données contemporaines.

### 2.2.4 Mesure des performances

La mesure des performances pour les tâches de transcription est le taux d'erreur après alignement entre la sortie du système et la transcription manuelle au format STM (Segment Time-Mark), toutes deux normalisées. Les règles de normalisation des transcriptions de référernces sont décrites dans l'annexe C.

La mesure des performances sera effectuée à l'aide des outils de mesure des performances (v2.0) distribuées lors de la campagne ESTER 1, soit

| | |
|---|---|
| `normalize-v0.55` | normalisation des STM/CTM, |
| `normalize.v2.0.dic` | dictionnaire d'équivalence orthographique, |
| `score-trs-v1.6` | calcul du taux d'erreur. |

L'outil `sclite` (version 2.2) sera utilisé pour l'alignement des soumissions avec les référence. Des mises à jour de ces outils, en particulier du dictionnaire d'équivalence orthographique, pourront être réalisées au cours de la phase de développement. La phase d'adjudication des résultats du test donnera lieu à une mise à jour du dictionnaire d'équivalence orthographique qui sera utilisé pour établir les résultats officiels de la campagne.

## 2.3 Extraction d'information

Les tâches d'extraction d'information visent à exploiter les transcriptions automatiques pour extraire des informations d'ordre linguistique. Dans le cadre de cette évaluation, nous considérerons la segmentation en unités syntaxiques et la détection des entités nommées.

Les tâches d'extraction d'information sont chacune divisées en deux sous-tâches, selon que le travail sera effectué sur une transcription de référence (*e.g.* SP-REF) ou sur une transcription automatique (*e.g.* SP-RAP). Afin de permettre aux participants ne disposant pas d'un système de transcription d'effectuer les tâches d'extraction d'information, le calendrier pour ces tâches sera décalé et les transcriptions d'un système de référence[5] seront rendues disponibles à ceux qui le souhaitent.

### 2.3.1 Segmentation en unités syntaxiques (SP)

Cette tâche exploratoire a pour but de segmenter un document en unités syntaxiques, ou phrases, comme défini dans les conventions de transcription (version 1.22). Sera considéré comme unité syntaxique toute entité syntaxique comprise entre deux marqueurs de ponctuation forts dans la transcription de référence, où les marqueurs de ponctuations forts sont

- le point (.),
- le point d'interrogation ( ?),
- le point d'exclamation ( !).

Les conventions d'annotation[6] spécifient que les marques de ponctuation correspondent aux pauses dans le signal liées à des frontières syntaxiques, les marques de ponctuations fortes étant liées à des balises de segments dans la transcrition de référence (changement de segment après la ponctuation).

L'objectif du système de segmentation en unités syntaxiques est de détecter les frontières d'unités syntaxiques dans la transcription automatique ou dans une transcription manuelle exempte de toute marque de ponctuation[7]. Il est à noter qu'aucune caractérisation du type des marques de ponctuation (affirmatif, interrogatif, etc) n'est requise. À partir d'une transcription sans ponctuation, le système doit produire une transcription contenant les marques de ponctuations fortes selon les formats spécifés en annexe de ce document.

Les données disponibles pour la campagne n'ayant pas fait l'objet d'une convention très stricte concernant l'annotation des frontières des groupes syntaxiques, cette tâche est à considérer comme une première tentative visant à définir de telles conventions. Nous rappelons que, en tant que tâche exploratoire, aucun classement officiel des systèmes ne sera communiqué en dehors du cadre des participants.

### 2.3.2 Détection d'entités nommées (EN)

Cette tâche consiste à détecter dans les transcriptions orthographiques automatiques et manuelles les références directes d'entités nommées (personnes, lieux, etc.), de dates et de montants, puis de définir le type de ces différentes entités.

Les types considérés dans cette évaluation sont :

[5] Non encore déterminé.

[6] Manuel de transcription, version 1.22 (http ://http ://trans.sourceforge.net/en/transguidFR.php)

[7] et par conséquent de casse en début de phrase !

– personne
– fonction
– organisation
– lieu
– production humaine
– date et heure
– montant

Une précision complémentaire du sous-type n'est pas exclue dans les résultats soumis mais ne sera pas prise en compte pour les évaluations de la phase 1. Les sous-types considérés ainsi que le détail des conventions d'annotations des entités nommées sont consignés dans un document distribué aux participants.

Ces conventions diffèrent légèrement de celles ayant servies à l'annotation du corpus de la campagne ESTER 1, notamment dans la répartition du jeu d'étiquettes. Cependant, une adaptation des annotations ESTER 1 devraient les rendre utilisables pour l'apprentissage dans le cadre de cette nouvelle campagne. Les données d'apprentissage du corpus ESTER 2 ne sont pas aujourd'hui annotées en entités nommées. Les données de développement sont elles annotées en entités nommées selon les nouvelles conventions.

### 2.3.3 Mesure des performances

**Segmentation en unités syntaxiques.** La segmentation en unités syntaxiques est évaluée selon le taux d'erreur défini comme

$$\text{error} = \frac{\#\text{ insertions} + \#\text{ suppressions}}{\#\text{ frontières de référence}} \tag{3}$$

où les nombres d'insertions (# insertions) et de suppressions (# suppressions) sont déterminés après alignement des mots (et marques de ponctuations) de la transcription de références avec la sortie du système.

Des outils de mesure de performance pour la tâche de segmentation en phrase seront distribués après validation du cahier des charges pour cette tâche.

**Détection d'entités nommées.** La principale mesure de performance sera une variante pondérée du *slot error rate* ($SER$) défini par Makhoul[8]. Le principe de base est de fournir un taux d'erreur sur l'ensemble des entités de référence. La base de calcul est donc le nombre d'entités, et non pas le nombre de mots contenus dans les entités de référence.

Pour cette variante du *slot error rate*, on peut distinguer trois types d'erreurs :

– $I$ : les insertions, qui sont des entités détectées dans l'hypothèse et qui n'ont aucun mot commun avec une entité de référence.
– $D$ : les délétions, qui sont des entités de référence totalement manquées par le système.

[8] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, Performance measures for information extraction, in Proceedings of DARPA Broadcast News Workshop, February 1999.

– $S$ : les entités de référence substituées, correspondant de façon incorrecte à des entités d'hypothèse.

L'ensemble $S$ lui-même se décompose en plusieurs catégories d'erreurs : les erreurs simples (extension correcte avec type incorrecte - $T$, ou extension incorrecte avec type correct - $E$), et les erreurs complexes (type et extension incorrects). Ces dernières se décomposent encore en : une entités de référence correspondant à une seule entité d'hypothèse ($TE$ pour type + extension incorrectes), ou une entité de référence correspondant à plusieurs entités d'hypothèse ($M$ pour multiple).

Si on note $R$ l'ensemble des entités de référence, taux d'erreur $SER$ est alors défini par :

$$SER = \frac{\#I + \#D + \#S}{\#R}, \tag{4}$$

ou encore

$$SER = \frac{\#I + \#D + \#T + \#E + \#TE + \#M}{\#R}. \tag{5}$$

Dans le script de scoring utilisé, chaque type d'erreur sera pondéré par un coefficient .

Une version adaptée des outils de calcul du slot error rate utilisés lors de la tâche exploratoire EN de la campagne Ester 1 Phase 2 sera distribuée ultérieurement. Les adaptations concerneront principalement la prise en compte du nouveau jeu d'étiquette adopté dans les annotations des données de cette nouvelle campagne.

# 3 Ressources

Dans le cadre de la campagne Ester 2 Phase 1, un corpus d'apprentissage d'environ 100 heures d'émissions radiophoniques sera distribué, en complément des données de la campagne Ester 1 Phase 2. Un sous-ensemble du corpus composé de 6h est identifié comme corpus de développement. Les participants sont invités à utiliser ces données de développement pour la mise au point de leurs systèmes afin de permettre une comparaison sur une base commune dès la phase de développement. Un corpus de test de 6h, proche des données dites de développement, sera distribué au démarrage de la campagne de test (cf. calendrier ci-dessous). Ces ressources sont décrites plus en détails dans la suite de ce document.

L'accès aux données d'apprentissage Ester 2 Phase 1 sera soumis à la signature d'un contrat d'utilisation entre le participant et la DGA. Les participants désirant obtenir les données transcrites Ester 1 Phase 2 devront signer un contrat d'utilisation auprès de ELDA. Ces deux contrats stipulent en particulier que les participants soumettant des résultats dans au moins une des tâches lors de la campagne de test pourront conserver les données gratuitement à des fins de recherche à l'issue de la campagne.

Toute ressource antérieure au 01/01/2008 peut être utilisées dans la phase de développement. En particulier, les ressources dites "non transcrites" de la campagne ESTER 1 Phase 2 pourront être utilisées sans aucune restriction : à cette fin, le Laboratoire d'Informatique de l'Université du Mans (LIUM) met à disposition 13 heures de transcriptions manuelles de parties conversationnelles des données "non transcrites" d'ESTER 1 Phase 2, transcriptions développées dans le cadre du projet ANR Étude de la Parole Conversationnelle (EPAC).

Les systèmes utilisant des ressources autres que celles fournies pour la campagne devront identifier ces ressources et fournir, dans la mesure du possible, des résultats contrastifs illustrant l'apport de ces ressources.

## 3.1 Ressources acoustiques

**Corpus Ester 2.** Les ressources acoustiques distribuées dans le cadre de la campagne ESTER 2 Phase 1 sont constituées d'émissions radiophoniques transcrites manuellement. Les conventions d'annotation sont détaillées dans deux documents séparés intitulés

- ESTER 2, Transcription Détaillée et Enrichie, Conventions d'annotation[9] (version 0.1)
- ESTER 2, Entités Nommées, Dates, heures et montants, Conventions d'annotation[10] (version 0.1)

Les émissions enregistrées contiennent des émissions d'information, des dossiers liés à l'actualité du moment et des émissions plus conversationnelles. Les données de développement, dont la liste est donnée en annexe E, ont été enregistrées au mois de juillet 2007. Le tableau suivant récapitule la répartition de l'ensemble des données.

| | Corpus d'apprentissage | | Corpus de développement | |
|---|---|---|---|---|
| Source | date | heures transcrites | date | heures transcrites |
| France Inter | 99-02 | 26 h | Jui. 07 | 2 h |
| RFI | 00-02 | 69 h | Jui. 07 | 40 min. |
| Africa n°1 | 03 | 10 h | Jui. 07 | 2h20 |
| TVME (ex RTM) | - | - | Jui. 07 | 1h |
| Corpus EPAC | 03-04 | 13h | - | - |
| Total | | 118h | | 6h |

**Corpus Ester 1.** Un récapitulatif des données distribuées dans le cadre de la campagne ESTER 1 Phase 2 est donné dans le tableau ci-dessous. Pour plus de détails, voir le plan d'évaluation ESTER 1 Phase 2 (version 1.1)[11].

[9] http ://www.afcp-parole.org/ester/docs/Conventions_Transcription_ESTER2_v01.pdf
[10] http ://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf
[11] http ://www.afcp-parole.org/ester/docs/plan-phase2-1.1.pdf

| Source | Heures transcrites | Heures non-transcrites |
|---|---|---|
| France Inter | 37 | 337 |
| France Info | 12 | 643 |
| RFI | 27 | 445 |
| RTM | 22 | – |
| France Culture | 1 | 252 |
| Radio Classique | 1 | – |
| total | 100 | 1677 |

Les ressources du corpus EPAC proviennent des sources de la campagne Ester 1 : France Inter, RFI, France Culture, France Info. Il s'agit des transcriptions de 13 heures de parole conversationnelle extraites des données "non transcrites" de la campagne Ester 1 Phase 2. Elles seront distribuées directement par les partenaires du projet EPAC.

### 3.2 Ressources textuelles

Les participants souhaitant accéder aux données textuelles distribués dans le cadre de la campagne Ester 1 Phase 2 (corpus "Le Monde" 1987–2003) doivent signer un contrat d'utilisation avec ELDA.

La distribution du corpus "Le Monde" 2004–2006 est à l'étude.

### 3.3 Ressources lexicales

Aucune ressource lexicale n'est spécifiée. Il existe cependant des phonétiseurs libres qui permettent la phonétisation du corpus d'apprentissage et du lexique. En particulier, le Laboratoire Informatique d'Avignon met à la disposition des participants son phonétiseur[12].

## 4 Règles de participation

Les participants s'engagent à fournir une description du ou des système(s) utilisés en spécifiant clairement les ressources utilisées, les algorithmes et méthodes mis en œuvre ainsi que le temps de traitement et la taille mémoire nécessaires. Ils s'engagent à venir présenter les travaux effectués dans le cadre de la campagne lors des ateliers Ester.

Pour l'ensemble des tâches, les règles suivantes s'appliquent :

– l'origine du document (c-à-d la chaîne de radio correspondant à un enregistrement) ainsi que la tranche horaire de l'enregistrement est une information qui peut être utilisée. Cependant, les données de test pourront provenir d'une tranche horaire pour laquelle aucune donnée d'apprentissage n'est disponible.

[12] `http ://www.afcp-parole.org/ester/repository/LIA/lia_phon.v1.2.tar.gz`

- les données utilisées doivent respecter les contraintes décrites dans la section 3 du présent document.
- Pour chaque tâche, les participants soumettant plusieurs systèmes devront identifier un système principal qui servira pour établir le classement officiel des participants. Les autres soumissions seront considérées à titre de contrastes.
- les résultats retournés après la date de clôture du test (cf. calendrier) ne seront pas considérés dans la classification des systèmes.

Par ailleurs, quelques règles essentielles sont rappelées ici :

- Les données audio ne peuvent être examinées ou écoutées avant ou pendant le test.
- Les systèmes évalués ne peuvent être modifiés une fois le traitement commencé. Un système ne peut être testé qu'une seule fois.
- Le traitement des données doit être entièrement automatique. Le résultat de ce traitement ne peut en aucun cas être modifié. Les seules interventions manuelles autorisées sont limitées aux opérations de lancement des traitements, aux vérifications de bon fonctionnement et aux opérations de relance éventuelles en cas de problème informatique.
- Si plusieurs systèmes sont évalués pour une tâche, aucun résultat ne peut être examiné avant la fin du dernier traitement à soumettre. Un système et un seul doit être identifié comme système primaire.

# 5 Calendrier

| | |
|---|---|
| janvier 2008 | distribution des données d'apprentissage et de développement |
| février 2008 | distribution des outils de mesure des performances |
| mai/juin 2008 | atelier à mi-parcours |
| juillet 2008 | version finale du plan d'évaluation |
| octobre 2008 | distribution des données de test |
| +15 jours | date limite de soumission des tâches T et S |
| | distribution de transcriptions automatiques pour les tâches E |
| +30 jours | date limite de soumission des tâches E |
| +45 jours | fin de la phase d'adjudication |
| décembre 2008 | atelier de clôture de la campagne |

# 6 Contacts

Guillaume Gravier (AFCP), guillaume.gravier@irisa.fr, 02 99 84 72 39
Laura Chaubard (DGA), laura.chaubard@etca.fr, 01 42 31 97 59
Edouard Geoffrois (DGA), edouard.geoffrois@etca.fr, 01 42 31 96 68
Khalid Choukri (ELDA), choukri@elda.org, 01 43 13 33 33

# A Format et métriques pour la segmentation

## A.1 Suivi d'événements

Pour les tâches de détection d'événements (SES et SVL), les résultats seront retournés au format ETF (Event Tracking File). Chaque ligne de la soumission correspond à un segment et un événement, indiqués selon le format suivant

```
source A début durée type sous-type événement score décision
```

où la signification des champs est

- source : nom du fichier sans extension ni chemin.
- début : temps de début du segment, en secondes par rapport au début du fichier
- durée : durée du segment, en secondes
- type : type d'événement (spk pour la tâche SVL, sc pour la tâche SES). Ce champ n'est pas utilisé pour la mesure des performances.
- sous-type : pour la tâche SVL, le sous-type peut-être male ou female ou encore unknown. Ce champ n'est pas utilisé pour la mesure des performances.
- événement : pour la tâche SVL, les événements correspondent aux noms des locuteurs ; pour la tâche SES, les événements sont music et speech.
- score : score associé à la décision ; plus le score est élevé, plus la décision est sûre. Ce champ n'est pas directement utilisé pour l'évaluation des performance mais permettra d'établir des courbes DET pour une meilleure comparaison des systèmes sur l'ensemble des points de fonctionnement. Ce champ peut être remplacé par un tiret (-) si aucun score n'est disponible.
- décision : décision de présence (true) ou absence (false) de l'événement recherché. Si ce champ n'est pas renseigné, l'événement est réputé présent.

Les lignes débutant par un point virgule seront traitées comme des lignes de commentaires.

Pour une tâche (SES ou SVL), les résultats d'un système seront soumis sous la forme d'un unique fichier ETF (encodage iso-8859-1). Le fichier de soumission spécifie l'ensemble des segments qui contiennent les événements considérés.

## A.2 Segmentation et regroupement de locuteurs

Pour la tâche SRL, les résultats seront retournées au format MDTM (Meta Data Time-Mark). Chaque ligne du fichier identifie un segment selon le format suivant

```
source A début durée type confiance sous-type id
```

où la signification des champs est

- source : nom du fichier sans extension ni chemin.
- début : temps de début du segment, en secondes par rapport au début du fichier
- durée : durée du segment, en secondes

- type : type d'événement ('speaker')
- confiance : mesure de confiance associée à la décision, dans l'intervalle [0,1]. Ce champ est optionnel et peut-prendre la valeur NA lorsqu'il n'est pas spécifié.
- sous-type : sous-catégorie parmi 'adult_male', 'adult_female', 'child' ou 'unknown' (champ non utilisé)
- id : identifiant arbitraire de locuteur (par exemple, loc1, loc2, etc.)

Les lignes débutant par un point virgule seront traitées comme des lignes de commentaires.

Un seul fichier MDTM (encodage iso-8859-1) par système sera soumis. Pour des raisons pratiques et afin de favoriser les systèmes produisant des segmentations réalistes, le nombre total de segment par fichier considéré sera limité à 5 000 par soumission. Si une soumission contient plus de 5 000 segments, seuls les 5 000 premiers seront considérés.

Le taux d'erreur de classification est établi en cherchant le meilleur appariement entre les locuteurs de la segmentation de référence et les identifiants arbitraires de la soumission. Le taux d'erreur est ensuite calculé à partir de cet appariement par comptage du temps total de segments (in)correctement classifiés.

Les segments non transcrits dans la référence (publicité) seront éliminés pour la mesure des performances. Une tolérance de 0.25 secondes sera appliqués aux frontières des segments de référence afin de ne pas pénaliser un léger décalage des frontières.

# B Format et métrique pour la transcription

Le format de soumission des résultats pour les tâches de transcription est le format CTM (Conversation Time-Mark). Chaque ligne de la soumission correspond à un mot avec une spécification de temps et un identifiant de fichier, suivant la syntaxe

```
source A début durée mot confiance
```

où `source` correspond au nom du fichier (sans extension, sans chemin), `début` au temps de début du mot en secondes par rapport au début du fichier, `durée` à la durée du mot en secondes et `confiance` à une mesure de confiance normalisée dans l'intervalle [0,1]. Un seul fichier CTM (encodage iso-8859-1) contenant les mots pour l'ensemble des fichiers du corpus de test sera retourné par système. De plus, ce fichier doit être trié par ordre croissant selon les trois premières colonnes : les deux premières par ordre alphabétique, la troisième par ordre numérique. La commande Unix sort +0 -1 +1 -2 +2nb -3 permet d'effectuer ce tri. Pour plus de détails, voir la documentation du logiciel sctk 1.2c[13].

Les soumissions seront évaluées à l'aide du script `score-trs.v`*`1.6`* fourni dans le package d'évaluation. Celui-ci effectue des normalisations pour ne pas compter comme erreur des variantes orthographiques autorisées. Le dictionnaire de normalisation sera augmenté de nouvelles équivalences par les organisateurs pour prendre en compte les données de test.

L'alignement des soumissions à la transcription de référence se fait en deux temps : un premier alignement temporel permet d'affecter les mots (CTM) aux segments de la transcription de référence (au format STM), sur la base des temps des instants d'occurrences des mots. Dans un deuxième temps, un algorithme d'alignement dynamique est utilisé indépendamment pour chaque segment de la référence.

Certains phénomènes nécessitent un traitement particulier et sont optionnels dans la transcription. Dans ce cas, aucune erreur n'est comptée si le mot est absent de la transcription. Cependant, les mots optionnels sont pris en compte dans le calcul du nombre total de mots dans la transcription de référence. Les mots optionnels correspondent aux phénomènes suivants :

- mots partiellement prononcés : ces mots sont indiqués dans la référence en mettant entre parenthèse la partie manquante du mot. Un mot reconnu à la place d'un mot partiel sera considéré comme correct si la partie transcrite du mot correspond au début du mot inséré.
- mots d'origine étrangère autre que noms propres et noms couramment utilisés en franais (par exemple, sandwich)

Les segments de parole vérifiant les conditions suivantes (dans la transcription de référence) sont ignorés dans la mesure des performances :

- segment contenant de la parole superposée
- segment contenant plus d'un mot prononcé dans une langue autre que le franais, sans compter les noms propres, les acronymes et les mots couramment utilisés en franais

[13] http ://www.nist.gov/speech/tools.

– segment correspondant à de la publicité (non transcrit dans la référence)
– segment contenant plus de deux mots d'origine étrangère autre noms propres et noms couramment utilisés en franais

# C Règles de normalisation des transcriptions

Les règles de normalisations sont décrites en détail dans la documentation du package de scoring. Elles sont rappelées de manière synthétique ici :

- la casse n'est pas prise en compte (tout les mots sont en minuscule)
- la ponctuation est supprimée
- un espace est inséré après les apostrophes liées à des élisions (l'importance, quoiqu'il, lorsqu'on, jusqu'à, etc), l'apostrophe étant maintenue dans le constituant de gauche. Les prefixes pouvant donner lieu à élision sont données dans le script perl ci-dessous.
- les mots composés (séparation par un tiret) sont divisés en leur constituants, le tiret étant supprimé.
- les expressions numériques sont récrites sous forme littérale
- les sigles sont laissés dans leur forme compacte (séquence de sans espace ni point). Cependant, pour les sigles contenant des chiffres (comme les noms de routes et d'autoroutes), l'équivalence dans laquelle la partie numérique est développée sous forme littérale est produite.
- les mots d'hésitations (euh, hum, huhum, mm) sont remplacés par le symbole %hésitation (aucune erreur n'est comptée pour une hésitation non reconnue)
- la partie non prononcée des mots partiellement prononcés est supprimée (aucune erreur n'est comptée lorsqu'un mot partiellement prononcé n'est pas reconnu ; un mot dont le début correspond orthographiquement à la partie prononcée du mot partiellement prononcé sera compté comme correct à l'alignement)
- les mots mal prononcés (mais pas tronqués) sont laissés tels quels (forme du dictionnaire).
- les mots marqués à orthographe incertaine, qui ne représentent qu'une proportion très faible des données, ne reoivent pas de traitement particulier même s'ils sont des candidats privilégiés à des entrées dans le dictionnaire déquivalence.
- un dictionnaire d'équivalence (par exemple, événement et évènement, clé et clef ou encore Hong Kong et Hongkong) sera fourni aux participants peu avant la campagne ; toute forme graphique apparaissant dans le dictionnaire est récrite en une forme unifiée.

L'orthographe des mots dans la transcription de référence a été validée par *aspell* et sert d'orthographe de référence pour les mots concernés dans les transcriptions. Les variantes othographiques ou grammaticales attestées dans les dictionaires Larousse et Robert, ou dans le Grévisse, sont acceptées. Les variantes fréquemment rencontrées sur internet peuvent également être acceptées.

Les participants peuvent proposer des mises à jour du dictionnaire jusqu'au démarrage de l'évaluation. Après soumission des résultats, les organisateurs produiront une proposition de mise à jour du dictionnaire pour tenir compte des données de test et des soumissions. Les participants auront 48h pour réagir à cette proposition. Les modifications n'ayant pas fait l'objet d'opposition seront intégrées pour la mesure des résultats officiels.

# D Format et métrique pour l'extraction d'information

## D.1 Segmentation en phrase

Les résultats de la segmentation en unités syntaxiques seront fournis au format STM-SB (Segment Time Marked - Sentence Boundary) pour la tâche portant sur les transcriptions de références (SP-REF) et au format CTM-SB (Converstation Time Marked - Sentence Boundary) pour la tâche sur les transcriptions automatiques (SP-RAP).

Ces deux formats sont des extensions des formats STM et CTM définis précédemment pour lesquels le champ `mot` est étendu pour inclure ou non une marque de segmentation, selon

```
mot = lexie[--<S/>]
```

la présence du marqueur `--<S/>` indiquant une frontière syntaxique après le mot défini par `lexie`. Il est à noter que le champ `confiance` des fichiers CTM s'applique à la lexie et non à la frontière détectée.

## D.2 Détection des entités nommées

Pour la **tâche EN sur transcriptions de référence** (EN-ref), les résultats seront retournés au format STM-NE (Segment Time-Marked - Named Entities). Ce format correspond à une modification du format initial STM, utilisé pour fournir les transcriptions de référence. Chaque ligne du fichier identifie un segment selon le format suivant :

```
source canal locuteur début fin <type> transcription
```

où la signification des champs est :

- `source` : nom du fichier sans extension ni chemin,
- `canal` : canal audio, toujours égal à `'1'` dans le cas présent,
- `locuteur` nom du locuteur du segment, tel que donné dans les transcriptions de référence, ou `excluded_region` dans le cas de segment ignores, ou `inter_segment_gap` dans le cas de segments vides
- `début` : date de début du segment, en secondes par rapport au début du fichier,
- `fin` : date de fin du segment, en secondes par rapport au début du fichier,
- `<type>` : type du segment,
- `transcription` : transcription orthographique.

Le champ `transcription` respecte le format suivant :

```
mot1 [typeEN1 mot2 mot3 ] mot4 [typeEN2 mot5 mot6 mot7 ] mot8
```

où chaque `mot[1-6]` représente un mot de la transcription, `mot2 mot3` est une première entité de type `typeEN1` et `mot5 mot6 mot7` est une seconde entité de type `typeEN2`. Les signes de ponctuation ne sont pas représentés ici dans un souci de clarté, mais ne sont pas exclus, y compris à l'intérieur d'entités : ceux-ci seront supprimés lors de l'étape de normalisation.

Une entité ne doit pas être découpée sur plus d'un segment, et la soumission ne doit pas comporter de soumissions imbriquées (dans le cas contraire, elles seront ignorées).

Le `typeEN` doit figurer parmi la liste suivante : `pers org loc gsp fac prod time amount` .

Attention : la détection d'entités nommées ne doit pas modifier les mots de la transcription de référence.

Pour la **tâche EN sur transcriptions automatiques** (EN-rap), les résultats seront retournés au format CTM-NE (Conversation Time-Mark - Named Entities). Ce format est une adaptation du format CTM décrit dans le plan d'évaluation (section A). Il respecte le format suivant :

```
source canal début durée mot(--typeEN--idEN) confiance
```

où les champs `source canal début durée confiance` ne diffèrent pas du format CTM initial. Seul le champ `mot` est modifié, dans le cas où il correspond à une entité. Dans ce cas, il est remplacé par `mot--typeEN--idEN` où `typeEN` correspond au type de l'entité (parmi `pers org loc gsp fac prod time amount` ), et `idEN` correspond à un identifiant pour l'entité. L'identifiant doit être unique pour chaque instance d'entité, c'est-à-dire qu'il doit être différent pour deux références distinctes de la même entité (voir les fichiers d'exemples joints au package de scoring).

Dans les deux cas, une précision complémentaire du type `typeEN` (sous-type, métonymies) n'est pas exclue dans les résultats soumis mais ne sera pas prise en compte pour les évaluations de la phase 2. Elle doit alors respecter le format suivant :

```
type1.soustype1/type2.soustype2
```

où `soustype` vient préciser le `type`, et où le `'/'` sépare les deux types dans le cas d'une métonymie. Attention : pour la métonymie, seul le second type, `type2` , sera pris en compte.

Au niveau des références utilisées pour l'évaluation, les métonymies seront considérées comme des alternatives. Les précisions des sous-types ne seront pas pris en compte.

# E Liste des fichiers de développement

Le tableau ci-dessous dresse la liste des fichiers transcrits pour le corpus de développement de 6h de la phase 1.

| nom du fichier | durée transcrite |
|---|---|
| 20070608_0730_AFRICAN1.trs | 0 :16 :49 |
| 20070613_0730_AFRICAN1.trs | 0 :17 :59 |
| 20070614_0730_AFRICAN1.trs | 0 :15 :55 |
| 20070615_0730_AFRICAN1.trs | 0 :12 :54 |
| 20070618_0730_AFRICAN1.trs | 0 :14 :44 |
| 20070619_0730_AFRICAN1.trs | 0 :16 :51 |
| 20070625_0730_AFRICAN1.trs | 0 :16 :53 |
| 20070626_0730_AFRICAN1.trs | 0 :15 :21 |
| 20070628_0730_AFRICAN1.trs | 0 :15 :03 |
| 20070707_0700_RFI.trs | 0 :28 :52 |
| 20070710_0631_RFI.trs | 0 :16 :10 |
| 20070710_1900_FINTER.trs | 0 :19 :05 |
| 20070711_1900_FINTER.trs | 0 :18 :42 |
| 20070712_1900_FINTER.trs | 0 :18 :11 |
| 20070716_1207_FINTER.trs | 0 :20 :04 |
| 20070723_1923_FINTER.trs | 0 :36 :27 |
| 20070715_2043_TVME.trs | 0 :17 :40 |
| 20070716_2044_TVME.trs | 0 :15 :23 |
| 20070717_2044_TVME.trs | 0 :16 :27 |
| 20070718_2044_TVME.trs | 0 :14 :27 |