ICiMa
CHAIRE D'INNOVATION
CIRQUE ET MARIONNETTE

CNAC CENTRE NATIONAL DES ARTS DU CIRQUE

INSTITUT INTERNATIONAL DE LA MARIONNETTE

Cahier des charges

# Transcription d'entretiens audios à visée terminologique

Actualisé le 27 février 2019

Contact :
Stéphane Riou, ingénieur de recherche - termino-icima.institut@marionnette.com - 03 24 33 72 68

# 1. Contexte du chantier : ICiMa, axe 3 – TERMINOLOGIE MULTILINGUE DES ARTS DE LA MARIONNETTE ET DES ARTS DU CIRQUE

## 1.1. Contexte général

La chaire d'innovation cirque et marionnette (ICiMa) <http://icima.hypotheses.org> travaille sur un double projet de terminologie multilingue des arts du cirque et des arts de la marionnette dont la nécessité est liée à l'absence de lexique spécifique et d'outils de traduction assistée par ordinateur, à l'insuffisance des outils permettant de décrire les objets et les pratiques, à la nouveauté des référentiels de compétences métiers et la nécessité de nommer et de pouvoir décrire les nouvelles formes des arts de la marionnette et du cirque notamment issus de la création contemporaine qui multiplie les pratiques d'hybridation et de métissage, ainsi que les innovations techniques et esthétiques.

### 1.1.1. Absence des lexiques relatifs aux arts de la marionnette et aux arts du cirque dans les dictionnaires et outils de traduction assistée par ordinateur.

Cette absence rend aujourd'hui difficile voire impossible le travail des traducteurs et interprètes, voire celui des artistes et des chercheurs lors des tournées, colloques et autres rencontres ou projets internationaux. Or nos pratiques s'exportent bien et beaucoup, de même qu'elles ne cessent d'accueillir formes, idées et acteurs venus de l'étranger.

### 1.1.2. Insuffisance des outils permettant aux métiers du patrimoine, de la documentation et de la médiation de décrire les objets et pratiques relatifs à nos disciplines.

Faute de thesauri adaptés, musées et bibliothèques se trouvent contraints à l'imprécision, au lacunaire, voire au détournement de sens. En ne reflétant pas la diversité des pratiques, ceci contribue notamment à renforcer la pauvreté des imaginaires collectifs relatifs à nos arts véhiculés notamment par le biais d'Internet.

### 1.1.3. Formulation récente ou en cours des référentiels de compétences des métiers

La formulation récente ou en cours des référentiels de compétences des métiers relatifs à nos disciplines et de la mise en place des certifications et diplômes afférents (DNSP d'acteur-marionnettiste, projet de DMA de constructeur ; DNSP cirque, DE-Cirque, certificat en dramaturgie circassienne en partenariat entre l'ESAC de Bruxelles et le Cnac) ;

### 1.1.4. Nécessité de nommer

La nécessité de nommer – de la part des artistes, des pédagogues, comme de la critique, des chercheurs, du public, ou des institutions – les nouvelles formes de marionnettes ou d'agrès, de nouveaux gestes, de nouvelles esthétiques et dramaturgies, dont la création accompagne l'effervescence de la création contemporaine.

## 1.2. Enjeux

Ce chantier de recherche vise donc d'une part à établir un état des lieux et une analyse historiques et géographiques des champs sémantiques concernés afin de mettre en évidence leur diversité diachronique et synchronique, les contextes socioculturels dans lesquels s'inscrivent ces pratiques, ainsi que leurs implications anthropologiques.
D'autre part, il doit aboutir à la production d'outils et de référentiels pour la traduction d'ouvrages, l'interopérabilité des bases de données (dans le contexte du web sémantique), l'interprétariat simultané afin de permettre une meilleure intercompréhension des professionnels, des chercheurs et du public.

## 1.3. Projet de recherche

Cette recherche fait appel à des compétences en linguistique théorique et appliquée, théories du spectacle, pratique artistique, anthropologie et philosophie. Elle implique une équipe pluridisciplinaire (historiens du cirque et de la marionnette, linguistes, terminologues et spécialistes du traitement automatisé de la langue) et la consultation d'artistes et de professionnels de la marionnette et du cirque.

Après définition des outils et méthodes en concertation avec les laboratoires spécialisés en TAL et en terminologie, cette recherche débute par une enquête internationale visant à collecter le vocabulaire des praticiens, celui des chercheurs et historiens, mais aussi celui de la critique et du public. Celle-ci est complétée par l'exploration de corpus écrits (manuscrits, imprimés) ou audiovisuels (réalisation et exploitation d'entretiens avec des praticiens notamment).

L'analyse de ces données, collectées et organisées sous forme de base de données interopérable, donnera lieu à des travaux comparatistes mettant au jour de nouvelles connaissances historiques et ouvrant de nouvelles perspectives sur les réalités contemporaines (journées d'étude, mémoires de Master ou thèse de Doctorat, publications) mais aussi à la création d'outils et de référentiels (création de néologismes, mise à jour, développement et restructuration de thesauri, édition de glossaires, ontologies, rédaction de guides de bonnes pratiques à l'intention des traducteurs et des institutions du patrimoine) destinées à améliorer la description des pratiques des arts de la marionnette et arts du cirque, à permettre l'injection de ces champs sémantiques dans les outils de Traduction Assistée par Ordinateur, à en améliorer le référencement dans le web sémantique et à

faciliter l'intercompréhension entre les praticiens, les chercheurs et les publics de ces disciplines sur le plan international. À travers ces différentes innovations, il s'agit de contribuer au renouvellement des imaginaires collectifs sur les arts du cirque et arts de la marionnette ainsi qu'au développement de leurs publics.

**Le présent périmètre de la recherche concerne le discours des praticiens en contexte de création ou de transmission (discours adressé à d'autres praticiens ou à des apprenants) et/ou portant sur la création ou la transmission de leur pratique (lors d'entretiens réalisés au cours de cette recherche).**

## 1.4. Contexte spécifique : Traitement de transcriptions d'entretiens

Le projet de l'axe 3 « terminologie multilingue des arts de la marionnette et des arts du cirque » comporte plusieurs volets.

Le premier est un volet analytique visant à examiner l'état des pratiques langagières des praticiens du cirque et de la marionnette en contexte (corpus actuellement mobilisé : discours des praticiens s'adressant à d'autres praticiens du début du 19e siècle à aujourd'hui). Pour constituer ce corpus, nous devons croiser plusieurs approches :

- Traitement d'imprimés
- Traitement de transcriptions d'entretiens
- Traitement de transcriptions de manuscrits (notes de mise en scène, correspondance, journaux de bord etc.)
- Et, vraisemblablement, car des traces textuelles ou audiovisuelles n'existent pas toujours, des formulaires déclaratifs remplis par les praticiens.

Le second volet consiste en la production d'outils de référence : dictionnaires, aides à la traduction, thésaurus, guides de bonnes pratiques à l'intention des traducteurs, ou à l'intention des professionnels du patrimoine pour la description des documents et objets afférents à ces disciplines.

### 1.4.1. Traitement de transcriptions d'entretiens

C'est dans le contexte spécifique du traitement de transcriptions d'entretiens que ce cahier des charges est rédigé.

### 1.4.2. Descriptions du corpus d'entretiens audios

Le corpus à transcrire est un corpus d'interviews d'artistes au format audio.

# 2. Contraintes

## 2.1. Métadonnées et fichiers fournis

On fournira au prestataire un fichier XML/TEI par entretien

### 2.1.1. Fichier TRJS & fichier audio

Pour chaque entretien à traiter, un fichier audio exploitable par le logiciel TranscriberJS est fourni.
Celui-ci s'accompagne d'un fichier TRJS (fichier au format XML/TEI).

### 2.1.2. Metadata du fichier

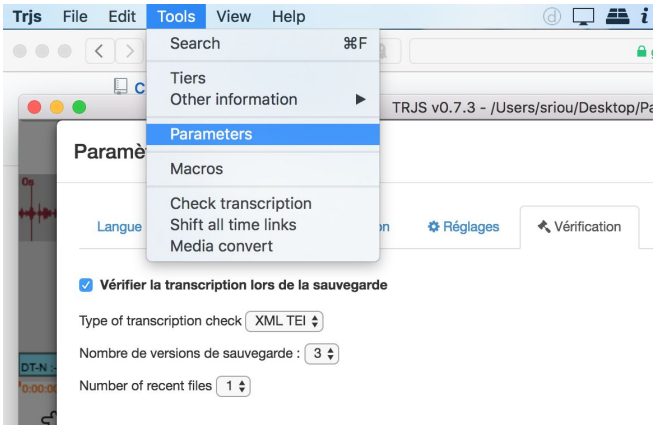
Le fichier TRJS associé au document à transcrire contiendra un header XML/TEI déjà prérempli.

### 2.1.3. Macros et balises associées

Les macro-commandes d'édition (raccourcis appelés dorénavant macros) sont des commandes qui permettent d'insérer des balises spécifiques dans les transcriptions.
Les macros et leurs descriptions sont spécifiées dans un document annexe. Il est possible que les macros doivent être entrées une première fois manuellement dans le logiciel par le transcripteur.
Les macros fournies ou décrites ne sont pas limitatives et le transcripteur est libre d'utiliser, en complément de celles-ci, ses propres macros s'il en éprouve l'utilité.
Toutefois, les macros ayant pour objectif un gain de productivité et un contrôle qualité, l'usage des macros fournies est **prioritaire** et surtout, et dans tous les cas, l'encodage lié aux balises insérées par les macros doit être scrupuleusement respecté. **En aucun cas le transcripteur ne pourra utiliser son propre balisage.**

## 2.2. Utilisation de TranscriberJS

Le Logiciel de transcription, d'édition et de visualisation de données et de corpus de langage oral à utiliser est TranscriberJS. La version à utiliser sera précisée avec les macros.
Téléchargement et documentation via ce lien : http://modyco.inist.fr/trjs/doku.php?id=start .
**Attention : Mettre les Paramètres de vérification de TRJS en XML TEI.**



# 3. Principes de transcription

## 3.1. Transcription fine & transcription basique

### 3.1.1. Transcription fine

La transcription fine est la transcription qui s'applique à toutes les parties du document associées aux objectifs de cette recherche. De façon non limitative **il s'agit essentiellement des passages abordant la pratique artistique, technique et/ou professionnelle, où l'on s'attend à trouver une plus forte densité de discours métier.**

### 3.1.2. Transcription basique

On entend par transcription basique la transcription des passages ne relevant pas d'une transcription fine.

## 3.2. Éditions des locuteurs

Les informations relatives aux locuteurs typiques seront fournies dans les méta-données.
S'il arrivait qu'un ou plusieurs locuteurs non-identifiés préalablement interviennent en milieu d'enregistrement, le transcripteur doit créer le nouveau locuteur dans les méta-données et affecter au nouveau locuteur les transcriptions relevant de celui-ci.

## 3.3. Transcription de l'audio

### 3.3.1. Type de transcription

Pour faciliter la lecture, c'est la transcription orthographique qui est choisie [En de rare cas, la transcription phonétique est utilisée quand la transcription orthographique est trop loin de ce qui est entendu --- valable pour la transcription fine].
Les débuts des énoncés ne sont pas marquées par des majuscules qui sont employées seulement pour les noms propres et les sigles courants.

Par souci d'optimisation du temps par rapport aux objectifs premiers de la présente recherche, **pour les passages en transcription basique, on se concentrera sur la transcription orthographique normée, dans laquelle on repérera à l'aide des macros**

**les rires et les soupirs, mais les autres phénomènes intéressant l'analyse conversationnelle n'ont pas à être marqués.**

#### 3.3.2. Tour de parole

Les tours de parole sont indiqués par le changement d'identification de locuteur (via la catégorie Participant [Person]).
Néanmoins un tour de parole qui excéderait 20 secondes devra être décomposé en unités plus petites à l'intérieur de ce tour de parole.
Les tours de parole doivent être indiqués aussi bien pour les passages en transcription fine que pour ceux en transcription basique.

## 3.4. Marquage des passages métier (donnant lieu à transcription fine)

Il sera demandé dans un tiers spécifique de marquer le document, ce qui correspond à une sorte de méta-balisage entre les passages sans importance directe pour le projet de recherche et ceux liés au projet. C'est le template "Fine" qui sera attribué dans le champ *Locuteur* pour les passages faisant l'objet d'un transcription dite fine (passages métier).

## 3.5. Marquage des identités, des lieux, des spectacles

Chacune de ces types d'entités nommées donnera lieu à un marquage spécifique à l'aide d'une balise (macro) afférente. Il y aura donc 4 types de macros liées à ce type de marquage.

#### 3.5.1. Lieu

#### 3.5.2. Date

Trois macros (et donc trois encodages) sont possibles pour marquer les dates en fonction qu'il s'agit d'une date complète, une date partielle avec le mois et l'année ou une date contenant seulement l'année.

#### 3.5.3. Identités

C'est les noms propres présent dans la transcription, qu'ils relèvent d'une personne réelle ou d'un personnage (ex. "Jacques Chesnais ou Guignol" sont des noms propres à marquer comme identité), ou encore d'une personne morale (collectif, compagnie, institution).

#### 3.5.4. Titre et nom de spectacles

Cf. liste et description des macros.

## 3.6. Transcription fine de l'audio (passage métier)

C'est également la transcription orthographique qui est choisie.

### 3.6.1. Doute, choix de transcription et utilisation de la macro "*incertain"*

Lorsque le transcripteur a un doute raisonnable, il doit utiliser la macro *incertain* spécifique aux hésitations du transcripteur (cf. liste et description des macros), et indiquer plusieurs interprétations possibles en séparant les termes par un pipe "|".

Dans de très rares cas, si la transcription orthographique est trop éloignée de ce qui est entendu, il faut utiliser la transcription phonétique.

### 3.6.2. Allongement du son (allongement vocalique)

Si l'articulation du son précédent a été notablement allongée, **:**, **::**, **:::** indiquent trois degrés. Ces points sont placés immédiatement à la suite du graphème qui transcrit le son allongé (« des: des:: »). Ces marqueurs d'hésitation sont appelés *allongement vocalique*.

### 3.6.3. Marques d'auto-correction, de changement de termes visés

Nous utiliserons dans nos transcriptions la macro associée à ces phénomènes d'élaborations non-abouties (qui permet de documenter notamment le terme visé) lorsque, dans le passage d'un discours métier et concernant un terme métier, le locuteur commence à prononcer un terme puis se reprend. S'il est impossible de trouver le terme visé, recopier exactement le contenu de la balise <sic> dans la balise <corr>.

Exemple : des <choice><sic>pou</sic><corr>poupées</corr></choice> des marionnettes

Cf. Liste et description des macros.

### 3.6.4. Chevauchements et interruptions « vacantes »

Le chevauchement et les interruptions sont traités automatiquement via le time-code associé aux tours de parole de chaque locuteur. Dans les cas d'interruptions ou de chevauchements n'ayant pour visée que de maintenir la dynamique conversationnelle (régulateurs de conversation du type : hum hum, oui oui, ok…, je vois), ces chevauchements et interruptions ne seront pas transcrits SAUF s'ils sont situés dans un passage de transcription fine ET s'ils témoignent de perturbations, constructions ou élaborations conjointes du discours métier ou de la définition d'un terme.

### 3.6.5. Pause courte & Pause longue

**Toute pause de moins de deux secondes sera considérée comme une pause courte au-delà de deux secondes, cela sera une pause longue.**

Exemple : ### pour indiquer une pause longue (supérieure à 2 secondes), # pour la pause courte.

#### 3.6.6. Indications gestuelles, mouvements, rires

Les indications gestuelles et les mouvements, les rires ainsi que les indications contextuelles informatives pour la compréhension de l'énoncé seront marquées à l'aide de la macro adéquate.
Cf. liste et description des macros.

#### 3.6.7. Marquage de termes métiers

Les termes ou énoncés relevant du discours métier seront marqués via une macro (dans un tiers spécifique). Cf. liste et description des macros.
En cas de doute, on préfèrera une signalisation excessive (on choisit le risque du bruit et non le silence).

# 4. Résultats attendus

## 4.1. Fichier .trsj

Le fichier .trjs qui est un fichier XML/TEI (et qui pourra posséder une autre extension) est le fichier fourni en début de transcription et qui contient les métadata relatives à l'enregistrement.
C'est ce fichier, augmenté du travail du transcripteur, qui doit être fourni en fin de transcription. Il doit être conforme aux normes indiquées dans ce cahier des charges et doit bien sûr être lié avec le document audio source.

## 4.2. Marge et taux d'erreur

#### 4.2.1. Repérage des passages métiers

Un taux d'erreur maximum de 7% par document transcrit, calculé par transcription et sur l'ensemble de la transcription, pour le repérage des passages métiers (correspondants à une transcription dite fine) sera acceptée.
Autrement dit, pour une transcription d'une heure, l'erreur de marquage des passages métiers qui correspond à la transcription fine du document ne pourra dépasser ce seuil.

#### 4.2.2. Erreurs dans la transcription

Le taux d'erreur est identique à celui donné au 4.2.1

#### 4.2.3. Erreurs dans la transcription fine

##### 4.2.3.1. Erreur de transcriptions

Le taux d'erreur est identique à celui donné au 4.2.1

##### 4.2.3.2. Erreur de repérage

Le taux d'erreur est identique à celui donné au 4.2.1

## 4.3. Confidentialité, sauvegarde et suppression des données

### 4.3.1. Confidentialité

Les documents et données échangées dans le cadre de cette recherche sont protégées par le droit d'auteur et par le droit à l'image. Elles ne peuvent pour l'instant être diffusées ni partagées en dehors du contexte précis de cette prestation et de cette recherche.

### 4.3.2. Sauvegarde

Le transcripteur est responsable de la sauvegarde de son travail en cours et achevé jusqu'à confirmation de la réception de la commande par la chaire ICiMa. Cette réception se fera via un courrier électronique dont l'objet comportera la mention "[ICiMa : réception de commande]".
Cette sauvegarde doit être effectuée dans un environnement sécurisé de sorte à garantir la confidentialité des données (DDE, clef USB etc.)

### 4.3.3. Suppression des données

Une fois la commande réceptionnée par la chaire ICiMa via le courrier électronique dont l'objet portera la mention "[ICiMa : réception de commande]", le transcripteur est tenu de supprimer de son outil de travail et de ses sauvegardes les fichiers relatifs à ce chantier (fichier(s) audio ; fichiers xm/tei).

### 4.3.4. Protocole d'échange

Les échanges de fichiers se feront uniquement par le moyen indiqué par la chaire ICiMa au moment du premier envoi.