**Formats pour les transcriptions de corpus de langage oral**

L’échange et la diffusion de transcriptions de langage oral, usages qui se généralisent et deviennent courants en linguistique et dans d’autres disciplines, a pour conséquence de rendre nécessaire aujourd’hui la création de formats et conventions communes au plus grand nombre d’utilisateurs possibles.

L’idée de format commun recouvre deux notions complémentaires :

* Disposer d’un format informatique normalisé et reconnu permettant de servir pour le stockage et la diffusion des données. Ce format doit être versatile et efficace d’un point de informatique ;
* Disposer de normes de présentation et de visualisation des marques linguistiques et des conventions de transcription pour permettre un accès facile à des corpus d’origines diverses et une bonne lisibilité de ceux-ci.

Politique générale pour les transcriptions de corpus

Plusieurs groupes de travail existent visant à arriver à ces formats communs existent aujourd’hui, notamment IRCOM (France), le groupe ISO/DIN européen, l’Equipex Ortolang. Ces groupes sont de nature différentes mais désirent travailler ensemble.

Politique de IRCOM

Deux groupes de travail d’IRCOM (consortium de la TGIR HUMA-NUM) ont travaillé sur ces questions : le Groupe 1 pour les normes de présentation et les données à coder, le Groupe 2 pour les codages informatiques.

Politique d’Ortolang

Dans le cadre d’Ortolang (cf. [www.ortolang.fr](http://www.ortolang.fr)) nous sommes amenés à mettre à la disposition de la communauté scientifique des corpus d’origines variées et qui sont susceptibles d’utilisations très diverses. Nous désirons donner un accès immédiat, visuel et sonore, aux corpus déposés, en utilisant la même interface pour tous les corpus. Ces corpus seront disponibles dans toute une variété de formats.

La politique d’Ortolang dans le choix des formats disponibles est de correspondre aux formats qui y sont déposés : les formats envisageables sont donc Transcriber, Praat, Clan, Elan, etc. sans limitation à priori, en restant dans ce qui est réalisable techniquement. Les représentations visuelles possibles sont donc aussi, potentiellement, celles des corpus qui seront déposés. Le format informatique sous-jacent doit être réalisé sur la base de la TEI et les conventions de transcription en accord avec les GT d’IRCOM.

Les membres d’IRCOM et d’Ortolang souhaiteraient initier une réunion de travail autour de ces thèmes pour mettre en commun les avancées des différents groupes.

Ce document est accompagné de :

* une proposition « groupe ISO/DIN » de format TEI
* une proposition « Ortolang » de format TEI pour les formats CLAN et TRANSCRIBER
* une proposition « Ortolang » pour la représentation graphique des transcriptions

**Conventions d’affichage pour la visualisation de corpus**

Pour la transcription de corpus oraux il est nécessaire de distinguer un format interne commun pour tous les corpus, et une représentation externe pas nécessairement commune. Pour Ortolang, nous avons créé un outil permettant la conversion de fichiers Chat et Transcriber vers un format TEI commun pour les corpus oraux, ainsi que la conversion dans le sens inverse, le but étant de réutiliser ce format commun à des fins de visualisation, d’interrogation de corpus et de conversion de formats. Pour la représentation, nous sommes au travail pour proposer un format unique pour les différentes sources, en respectant les besoins des corpus que nous avons à traiter.

Dans les corpus dont nous disposons, les codages pour certains phénomènes ne sont évidemment pas les mêmes. Pour plus de clarté lors de la visualisation et éventuellement l’édition de ces corpus il est nécessaire de se mettre d’accord sur des conventions de représentation commune des phénomènes présents dans les transcriptions.

A partir d’une liste de conventions qui nous a été proposée par IRCOM GT1 (cf. tableau) pour certains phénomènes qui sont codés dans les différents corpus à traiter mais pas tous, nous avons commencé à réfléchir à la mise en place de ces différents codages pour la visualisation des corpus. Les notes qui suivent présentent les conventions que nous proposons pour les phénomènes qu’on doit visualiser clairement.

**1) Mots orthographiques, transcription $**

*- Sigles, mots composés, lettres, acronymes*

Convention proposée : $ ... S/$

Exemple : $ ASSEDIC S/$

*-Toponymes*

Convention proposée : $ ... T/$

Exemple : $ Londres T/$

*-Titres*

Convention proposée : $ ... 0/$

Exemple : $ Le grand bleu 0/$

*-Patronymes*

Convention proposée : $ ... N/$

Exemple : $ Jean-Claude N/$

*- Incompréhensibles, indistinct*

Convention proposée : deux cas possibles:

- Le transcripteur croit distinguer le mot qui a été prononcé : $ ... ?/$

Exemple : $ train ?/$

- On n’a aucune idée de ce qui a été prononcé : $ X ?/$

Exemple : $ X ?/$

*- Plusieurs possibilités/alternance codique/mots étrangers*

Convention proposée : $ ..., ..., ... X/$

Exemple :

$ business, en X/$

$ des, les X/$

*- Parties non-transcrites*

Convention proposée : $ xxx /$

*- Amorces*

Convention proposée : ...--

Exemple : fabri--

**2) Mots phonologiques/phonétiques []**

*- API ou Sampa*

Convention proposée :

Deux cas :

- transcription orthographique et phonétique :

[mot, api ou sampa]

Exemple : [aéroport, aReopOR]

- transcription phonétique directe

[ ?, api ou sampa]

Exemple : [ aReopOR]

*- Syllabes, tons, intonation*

Convention proposée

[mot, transcription]

**3) Enoncés \***

*- Bruits*

Convention proposée : \* ... B/\*

Exemple : \* la porte claque B/\*

*- Variantes paralinguistiques (rires, pleurs, toux...)*

Convention proposée : \* ... P/\*

Exemple : \* rires P/\*

**4) Pauses et terminateurs #**

*- Pauses*

Conventions proposées :

3 cas :

- Pause courte : #

- Pause longue : ###

- Pause chronométrée : #2.3

*- Terminateurs (contour global)*

Convention proposée : #...

Exemples : #. # ? # !? ...

**5) Paralinguistique {}**

*- Hésitation*

Convention proposée : { ... H}­

*- Reprise/reformulation*

Convention proposée : { ... R}­

Exemple : ces {ce R} tableau

*- Contexte*

Convention proposée : { ... C}­

*- Explication*

Convention proposée : { ... E}­

**6) Syntaxe =**

*- Erreurs*

Convention proposée : =...

Exemple : la=le

*- Omissions*

Convention proposée : =0

Exemple : je veux voir =0 spectacle

*- Liaisons particulières*

Convention proposée : =...=

Exemple : donne =z= en

**7) Elision ()**

*- Elision*

Convention proposée : ...() ou ...(...)

Exemple : i(l) y a d(é)jà, d()jà

|  |  |  |
| --- | --- | --- |
| **Phénomène** | **PROPOSITION DE CONVENTION** | **Exemple** |
| Acronyme & sigle | $ ... S/$ | L1: il s'inscrit aux $ASSEDIC S/$ |
| Alternance de code | $ commentaire facultatif X/$ |  |
| Amorces de mots | ...- | L1: c'est une courti- |
| Bruits | \* type-du-bruit /\* | L1: je pense pas \*porte qui claque /\* que ça change |
| Chevauchements de parole | <id\_loc ... > | L1: c’est une retraite qui est financé par <L1 ces personnes-là> <L2 quand il y a> L2: plus de 2 millions de personnes |
| Chuchotant | $... C/$ | L1: va la mettre $sur l'étagère C/$ |
| Elision | (…) | L1: i(l) y a d(é)jà |
| En riant | $ ... R/$ | L1: non c'est pas vrai $me dit pas ça R/$ |
| Hésitation | liste close, lexicalisée | L1: non euh pas question |
| Lettres, mots épelés, sigles | $x x x x /$ | L1: mon nom c'est $G H I O /$ |
| Liaisons particulières | =...= | L1: donne moi =z= en |
| Mots en langue étrangère | [transcription, Sampa] | L1: on va au [B&B, biEnbi] |
| Mots indistincts (incompréhensibles, inaudibles, inconnu…), | ? | L1: fais moi passer le ? |
| Multi transcriptions, indécisions | /alt1, …, altn/ | L1: /des, les/ acides |
| Onomatopées | liste close, lexicalisée | L1: oops pardon |
| Parties enregistrées mais non transcrites | $ commentaire facultatif X/$ | $ commentaire facultatif X/$ |
| Patronymes | $ ... P/$ | L1: son nom c'est $Noël P/$ |
| Pauses | # | # |
| Prononciations particulières | [transcription, Sampa] | L1: je vais à l'[aéroport, aReopOR] |
| Rire en même parlant | $ … R/$ | L1: $pas possible R/$ |
| Titres | $... 0/$ | L1: $Le Grand Meaulnes 0/$ |
| Toponymes | $... T/$ | L1: on était à $Londres T/$ |
| Transcription phonétique directe | [?, sampa] | L1: un bruit genre [?, klo~g] |
|  |  |  |
| **Phénomènes non pris en compte dans la convention** | | |
| Abréviations |  |  |
| Accords non standards |  |  |
| Allongement |  |  |
| Autocorrection |  |  |
| Citation, texte lu |  |  |
| Conditions acoustiques |  |  |
| Conversation en parallèle |  |  |
| Discours rapporté |  |  |
| Léxèmes non standard lapsus | cf prononciation particulières |  |
| Négation |  |  |
| Nombres |  |  |
| Schwa |  |  |
| Sigle |  |  |
| Silence | cf. pauses |  |
| Syllabes inaudibles (nombre de syllabes (cf. ICAR)) | NON |  |
| Tour de parole emboité |  |  |
| Tours de parole |  |  |
| Transcription incertaine | NON |  |
| Troncation de mots |  |  |
| Variantes morphologiques indécidables | NON, cf. indécision |  |