



UMR 7309 • CNRS • AMU  
**PAROLE ET  
LANGAGE**

ANR  
AGENCE  
NATIONALE  
DE LA  
RECHERCHE

cnrs

Aix\*Marseille  
université

# Élaboration des guides d'annotation pour la transcription des corpus multimodaux à quatre mains

Benjamin Holt + Marion Tellier



# Le schéma d'annotation: définition

- Schéma : représentation schématique des éléments à annoter et leur hiérarchie
- A définir en fonction du corpus et de la question de recherche car pas de schéma universel « passe-partout ») (Kendon, 2004; Calbris 2011).
- Mettre le schéma à l'épreuve du corpus: va-et-vient entre le schéma d'annotation et les données (pour réajuster)
- Ce n'est pas échec de modifier son schéma en cours de recherche !

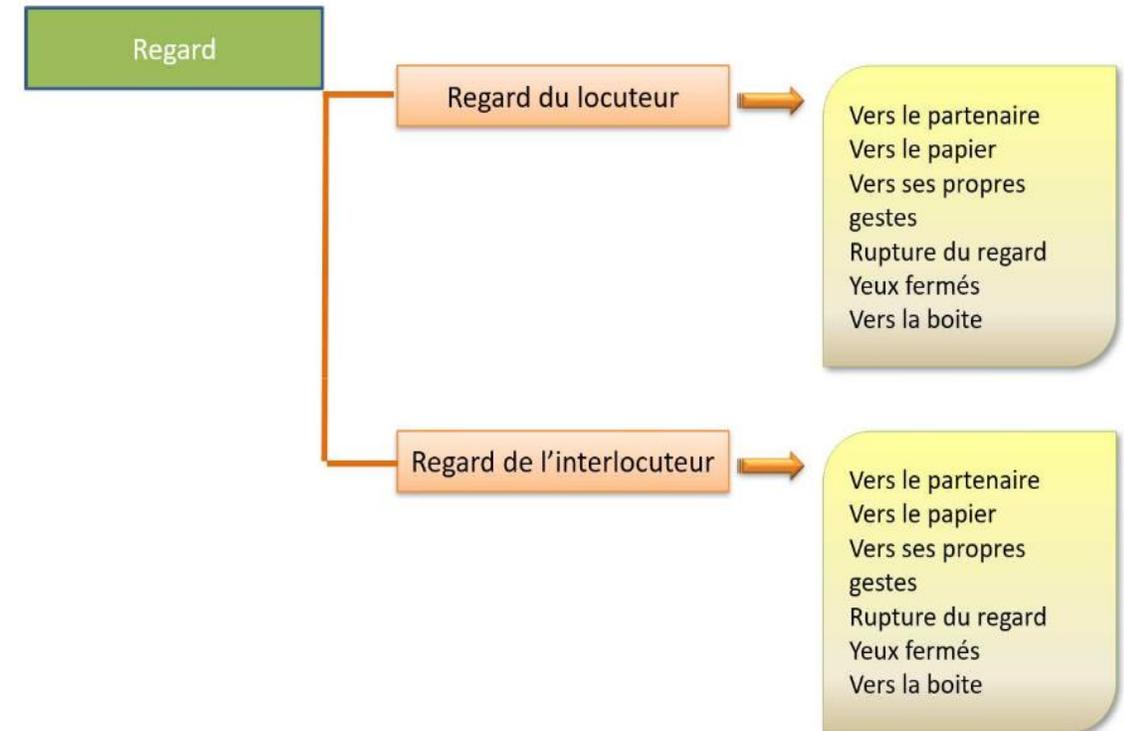
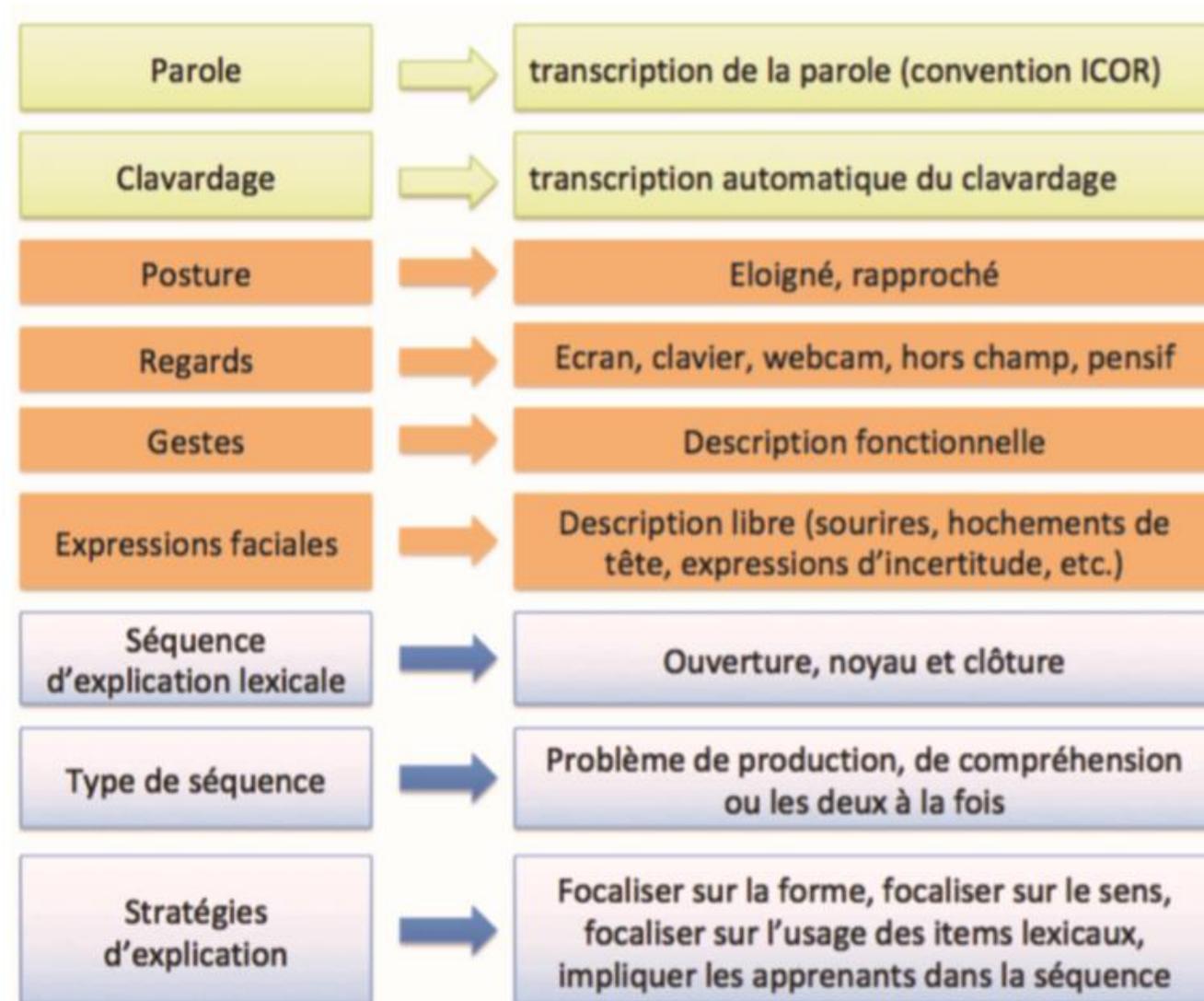


Figure 3. Schéma d'annotation



# Exemple d'un schéma d'annotation (thèse Holt)



# Exemple d'un schéma d'annotation sous ELAN

The screenshot displays the ELAN 5.7-FX software interface for a file named 'Victor\_S1.eaf'. The interface includes a menu bar (Fichier, Edition, Annotation, Acteur, Type, Rechercher, Affichage, Options, Fenêtre, Aide) and a toolbar with various editing and playback controls. The main window is divided into several sections:

- Video View:** Shows a live video of a man wearing a headset, with a smaller inset showing a stylized drawing of his face.
- Text View:** Displays the transcript of the audio. The selected text is 'grève 1', and the 'Explication lexicale' (lexical explanation) pane shows associated terms: 'contrôle · manifestation · grève 2 · mariage pour tous · 35 heures · congés payés · nu-pieds · paradoxe · organisation · gouvernement · pratiqué (supprimée) · pause café'.
- Waveform:** A black waveform representing the audio signal, with a blue selection box highlighting a segment between 00:03:13.440 and 00:03:14.234.
- Annotation Tracks:** A series of horizontal tracks for different annotation types:
  - Victor:** [448]
  - Liam:** [399] with phonetic transcription [°euh°]
  - Chat (Victor):** [2]
  - type de geste:** [27]
  - Explication lexicale:** [13]
  - Phase:** [34] with the annotation 'noyau grève 1'.
  - commentaires:** [33]
- Lexical Explanation Pane:** A pop-up window showing a list of terms: 'iconique', 'métaphorique', 'emblème', 'battement', 'deictique', and 'non identifi...'.



# Mauvais schémas d'annotation

- Ecart entre questions de recherche et schéma d'annotation
- Trop de catégories
  - plus il y a de catégories, plus c'est compliqué (on ne définit pas toujours bien la différence entre catégories)
- Catégories mal définis ou mal adaptées aux données
  - Un même phénomène peut appartenir à différentes catégories (frontières poreuses)
- Réutilisation difficile ou improbable



# Le guide d'annotation: définition

- On donne les outils nécessaires pour annoter les données
- C'est le manuel d'utilisation du schéma
- Doit pouvoir être utilisé par quelqu'un d'extérieur à la recherche (néophyte)
- Explications rédigées, exemples et photos si nécessaire
- Voir les guides mis en ligne par CORLI
- <https://corli.huma-num.fr/guides-annotation>



# Guides d'annotation en ligne

## Consortium CORpus, Langues et Interactions

[ACCUEIL](#)[PARTICIPER ▾](#)[ACTIONS ▾](#)[FORMATIONS](#)[BONNES PRATIQUES ▾](#)[RESSOURCES ▾](#)[GLOSSAIRE](#)[CONTACT ▾](#)

### Guides d'annotation

**Coordination:** Marie-Anne Sallandre

Toute entreprise d'annotation devrait s'accompagner d'un guide d'annotation détaillant les choix et les annotations posées sur les corpus. Néanmoins, il s'avère que le guide d'annotation est souvent ce que l'on rédige en dernier, et faute d'une version *\*parfaitement finalisée\**, on tarde souvent à le mettre en ligne.

Lorsqu'on annote un corpus et qu'on se lance dans l'écriture d'un guide d'annotation, on récupère les guides disponibles en ligne; il serait donc pertinent de pointer vers les guides d'annotation existants depuis le site de CORLI, et c'est l'objectif de la présente rubrique.

Un guide d'annotation est une publication acceptée par l'HCERES. Le site pointe donc vers les liens HAL de ces guides et propose également des mots clés (e.g. type de discours: LSF, écrits scolaires, projet, logiciel, état d'avancement + descriptif libre optionnel etc.).

### GUIDES D'ANNOTATION

- Balvet A., Barque L., Condette M.-H., Haas P., Huyghe R., Jugnet A., Marín R. et Merlo A. Le projet Nomage Description des noms déverbaux du français, 2009. ([pdf](#))
- Balvet A., Barque L., Condette M.-H., Haas P., Huyghe R., Jugnet A., Marín R. et Merlo A. Guide pour l'annotation des noms déverbaux, 2009. ([pdf](#))
- Benzitoun C. et Bérard L. Manuel d'annotation pour les corpus du projet PERCEO ([pdf](#))
- Bianchini C., Chèvrefils L., Danet C., Doan P., Rébulard M., et al. Guide d'annotation pour les polices de caractères Typannot dans le cadre du projet Typannot-Graphé. [Rapport de recherche] 1, Université de Rouen Normandie; ESAD Amiens; Université de Poitiers. pp.1-47, 2018. ([hal-01965272](#))
- Bigi B. Enriched Orthographic Transcription : SPPAS convention, 2017. ([pdf](#))
- Boutet D., Jégo J.-F., Meyrueis V. POLIMOD Pipeline: tutorial: step-by-step tutorial: Motion Capture, Visualization & Data Analysis for gesture studies. [Technical Report] Université de Rouen, Université Paris 8, Moscow State Linguistic University, 2018. ([hal-01950615](#))



# Raisons pour faire un guide d'annotation

- Être constant dans son annotation
- Permettre au même chercheur d'annoter d'autres données sur d'autres corpus, avec la même constance
- Permettre à d'autres chercheurs d'utiliser notre guide d'annotation
- Permettre le travail en équipe (plusieurs chercheurs utilisent les mêmes données)
- Contre-codage
- Aider le lecteur à bien comprendre ce qui a été annoté



# Exemples de guides d'annotation

## Comment annoter ?

On annote un geste depuis la préparation du geste (les mains commencent à bouger, l'image devient floue) jusqu'à la rétractation du geste (retour des mains en position de repos ou immobilité) ou l'enchaînement sur un autre geste. Voir Figure 15 (Le geste commence alors que la main est suspendue immobile depuis un moment).



Repos

Préparation

Stroke

Rétraction

Figure 15. *Ségmentation du geste*

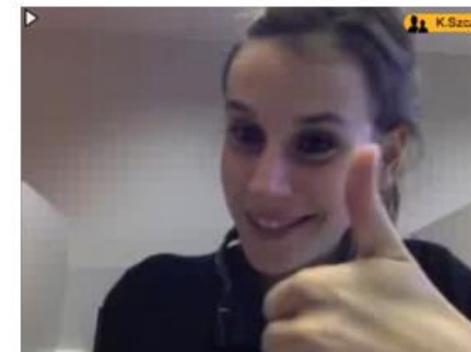
## Guide d'annotation GTT (Tellier et Stam)

## Emblème

Les emblèmes sont propres à chaque culture et peuvent être compris sans la parole. Ils sont conventionnalisés, ont un sens défini, et appartiennent à une communauté spécifique. Par exemple, le pouce vers le haut veut dire « okay » dans certains pays comme les Etats-Unis, et le geste de l'index tirant la peau du dessous de l'œil vers le bas pour dire « mon œil » (« je ne crois pas ») dans la culture française. (Tellier, 2008 ; Tellier, 2009).

Pour résumer : l'emblème est comme une expression idiomatique gestuelle. Sa forme est fixe et conventionnalisée. Les emblèmes appartiennent à des communautés socio-culturelles et ne sont pas forcément compris d'une communauté à une autre.

Dans cet exemple, l'interlocuteur produit un emblème qui sert à encourager l'apprenante. Elle tient sa main dans cette position quelques secondes après la fin de l'énonciation.



« [nickel] »

Guide d'annotation M2 (Holt)<sup>8</sup>



# Le contre codage

Une étape de validation des annotations par d'autres annotateurs

- Différentes appellations possibles (Royer et al. 2009)
  - procédures de validation du codage
  - accord « intercodeurs »
  - accord « interjuges »
  - « contre-codage »
  - « codeurs indépendants »
- Les diapos qui suivent sont inspirées de Tellier (2014) *Discours*



(1) À quel moment de la recherche proposer un contre-codage ?

Deux moments sont possibles

**A la fin du travail** : lorsque tout le corpus a été annoté

- Pour validation des annotations
- Echantillon à des annotateurs indépendants
- Cas le plus fréquent

**Au début du travail** : lorsque schéma et guide d'annotation sont conçus

- Pour les tester et faire émerger imprécisions et lacunes avant d'annoter tout le corpus
- Echantillon à des annotateurs indépendants
- Evoqué par Allwood *et al.* (2007) et Tellier *et al.* (2013)



(2) Quelle proportion du corpus contre-annoter ?

Pas de réponse unique

Dépend de la taille du corpus et de la complexité des annotations

15 à 20 % du corpus est une bonne base



### (3) Quels annotateurs (naïfs ou non, experts ou non)?



Codeur expert (linguiste) ou naïf ?



Codeur naïf : bien l'entraîner et lui donner un travail pas trop technique



Codeur expert : est habitué à l'annotation gestuelle, à la lecture d'un guide d'annotation et au logiciel



Mais l'expert peut être très tatillon sur certaines annotations voire accoutumé à des schémas d'annotation un peu différents !



Le non-expert se fie plus au schéma d'annotation et se pose peut-être moins de questions



(4) Que leur donner (schéma et guide d'annotation, étiquettes vides ou non) ?



Former les annotateurs (schéma et guide d'annotation + logiciel)



Méthode des étiquettes vides (on leur donne pré-segmenté et il leur suffit de choisir une étiquette)

(5) Que fait-on après le contre-codage ?

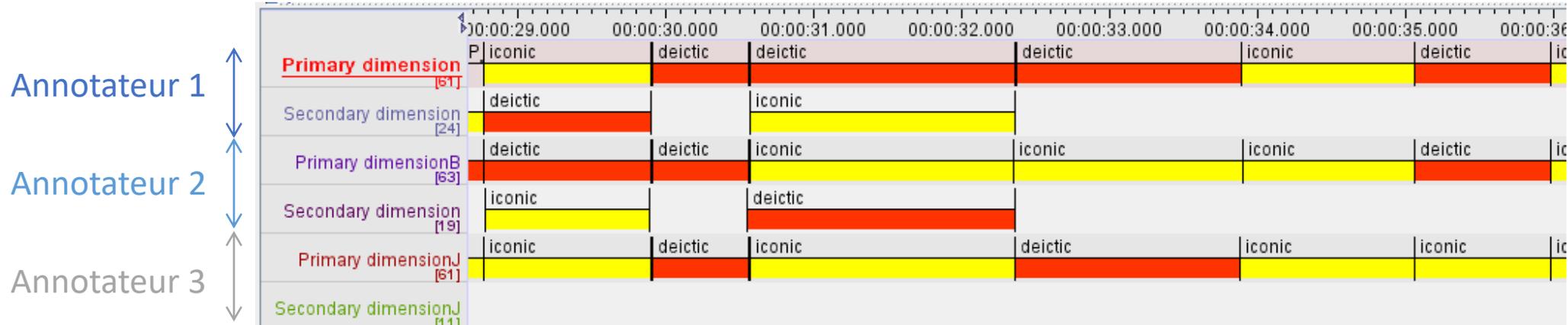
**Début d'étude:** rectifier le guide d'annotation si celui-ci manque de précision voire à modifier le schéma

**Fin d'étude:** permet de tester la fiabilité de l'annotation

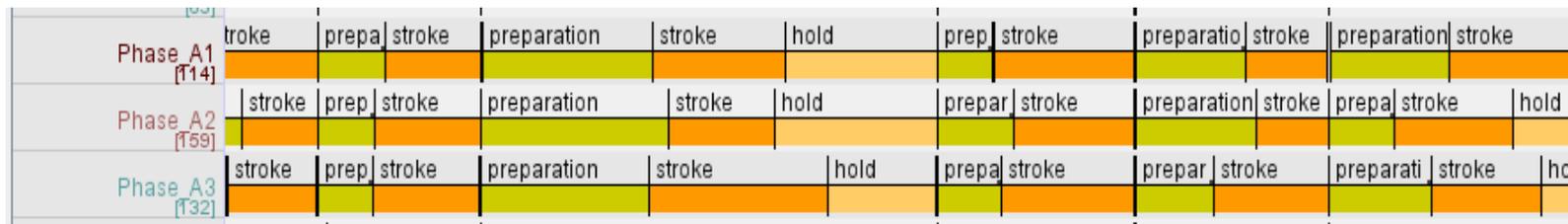
Dans un même fichier d'annotation, fusionner les codages des différents participants, pour regarder les points d'accord et de désaccord



# Accords et désaccords



Dimensions du geste



Phases du geste

# Quel calcul?

- Taux d'accord pour les unités gestuelles
- Logiciel Elan: kappa de Cohen (fonction *Compare annotators*).
- Seulement les segmentations temporelles et pas le contenu des annotations
- Étiquettes gestuelles = taux d'accord inter-juges très élevé :

| Juges    | Taux d'accord |
|----------|---------------|
| A1 et A2 | 0,948         |
| A1 et A3 | 0,863         |
| A2 et A3 | 0,904         |

« **Accord excellent** »  
Selon le barème de Landis et Koch (1977)

- Taux d'accord pour les dimensions primaires du geste
- Logiciel statistiques: kappa de Cohen et Fleiss

| Juges        | Taux d'accord |
|--------------|---------------|
| A1 et A2     | 0,520         |
| A1 et A3     | 0,347         |
| A2 et A3     | 0,281         |
| A1, A2 et A3 | 0,406         |

« **Accord moyen** »  
Selon le barème de Landis et Koch (1977)

