Les corpus de langage oral

QUESTIONS SCIENTIFIQUES, MÉTHODOLOGIQUES ET POLITIQUES

CHRISTOPHE PARISSE – CORLI & MODYCO – SEPT. 2021

Questions scientifiques

La langue orale et ses propriétés

La langue orale n'est pas la langue écrite et la langue écrite n'est pas une représentation de la langue orale

Références:

- **Blanche-Benveniste, C., Jeanjean, C. (1987)**. Le français parlé : Transcription et édition. *Et les ouvrages suivants*
- **Harris, R. (1990). On redefining linguistics**. In H. G. Davis & T. J. Taylor (Éds.), Redefining linguistics. London: Routledge. *Et tous les autres ouvrages*
- **Linell, P. (2005). The Written Language Bias in Linguistics** : Its Nature, Origins and Transformations (1er édition.). London ; New York: Routledge. *Voir aussi "Rethinking Language, Mind, and World Dialogically"*
- **Approche de l'analyse conversationnelle: Sachs, Schegloff, …** *voir Mondada, L. (2008). Contributions de la linguistique interactionnelle. Congrès Mondial de Linguistique Française. http://dx.doi.org/10.1051/cmlf08348*

Les corpus oraux ne sont pas une catégorie uniforme

Il existe de nombreux types de corpus oraux qui se différencient par:

- Les médias et les modalités (audios, videos)
- La cible scientifique: phonologie/phonétique, lexique/syntaxe, interaction, gestualité, pragmatique
- Les méthodes et les analyses (manipulation du signal, codages et transcriptions)

Quelques exemples

Guillaume, P. (1927). Les débuts de la phrase dans le langage de l'enfant. Journal de psychologie, 24, 3-25.

Page 16:
À partir du vingtième mois, la négation accompagne un grand nombre de verbes proprement dits : A po pas (je ne peux pas) ; Sais pas (20.4); A pas vu; A trouve pas (21.16). — Chez une autre enfant : A pas la même (12.19) (elle défend à son frère d'aller sur les genoux de sa mère); A pas télé ! A pas papa ! (Papa n'aura pas à téter) (13 m.); A pas tété papa (14 m.).

https://gallica.bnf.fr/ark:/12148/bpt6k127910q
[Archives de la parole]. , Fragment du discours prononcé à l'inauguration des Archives de la Parole / Mr Liard, aut., participant

Quelques exemples

Katja Ploog – disponible sur le corpus de la parole (Cocoon)

https://doi.org/10.34847/cocoon.25cdaedb-2084-30a2-a5af-9f7bdb2bc711

"4 Burkinabe (D)" 1997. Français. 23 (speaker); 25 (speaker); 26 (speaker); 27 (speaker); 08 (participant); 17 (participant); 19 (participant); 20 (participant); 24 (participant); 00 (interviewer); Ploog, Katja (depositor); Ploog, Katja (researcher). Editeur(s): Université Bordeaux III. @fr

Quelques exemples

ESLO1: entretien 004

https://doi.org/10.34847/cocoon.0f7a8a33-0cbd-3327-bb4f-ae481ced5758

Blanc, Michel ; Biggs, Patricia

(création: 1969-04-08)

[fr] Identifiant du témoin : DM 95 Homme, 28 ans (né en 1941, à Orléans, Loiret) ; ingénieur administratif. Enregistrement de sa femme et de leur nouveau-né. Enregistré par John Ross, le 8 avril 1969, au domicile du témoin

Goodwin (Talkbank)

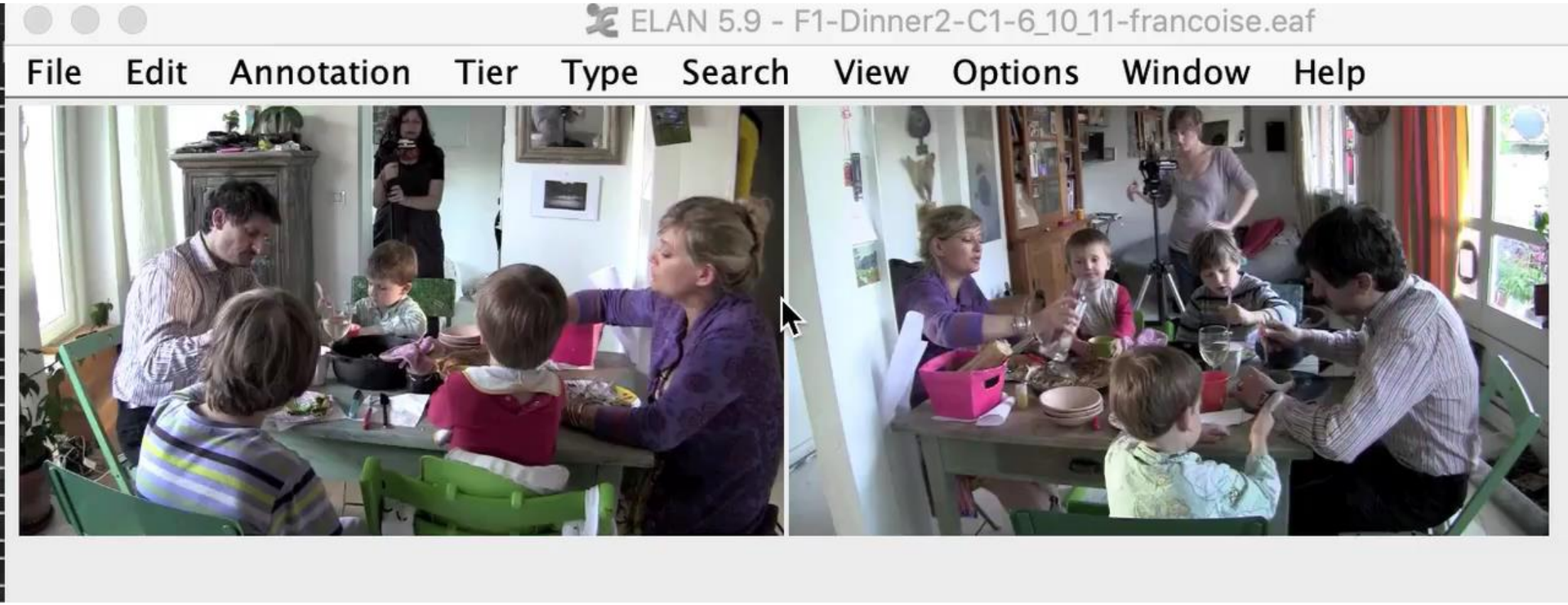
/Volumes/M2/Goodwin/cinco.cha

```
@UTF8
@Begin
@Languages: spa
@Participants:	C Carla Child, D Diana Child, M María Child
@Options:	CA
@ID:	spa|Goodwin|C|||||Child|||
@ID:	spa|Goodwin|D|||||Child|||
@ID:	spa|Goodwin|M|||||Child|||
@Media: cinco, video
*D:	&=enters_9 . •
*C:	hasta aquí. •
*D:	&=throws_bag &=steps_back . •
%com:	D threw into the wrong square, making route easier
*M:	oh, ⌜chin⌝. •
*D:	⌞&=steps_9 ⌟. •
*D:	&=steps_7 . •
*C:	⌜chiriona porque . •
*D:	⌞&=steps_6 &=steps_4 . •
*C:	⌜este es el cuatro . •
%gpx:	pushes D backwards with right hand
*D:	⌞&=falls_back . •
*C:	y tu vas en el cuatro &=hand_up . •
%com:	hand displays four fingers as C gazes directly at D
*C:	no vas ⌜en el quinto . •
%gpx:	holds up five fingers to display the error
*D:	⌞&=turns_down . •
%com:	turns to look at squares and feet
*C:	este es el quinto &=deictic_stomp . •
*D:	no . •
%gpx:	hands on hips in defense
09oct20[E|CHAT] 10
```

Movie - Sound

Anaé 2 ans 2 mois (corpus Colaje)

Corpus de diners parisiens



Langue des signes française

Projet ANR
Creagest

2007-2012

Une grande variété de recueils

Audio ou/et Vidéo:

- Enregistrement monologue en studio
- Conversation en studio
- Conversation en milieu naturel
- Conversation en milieu bruyant

Mêmes situations mais avec plusieurs sons, plusieurs vidéos

Une grande variété de codages

- Transcription orthographique (standard, normalisée, adaptée à l'oral)
- Phonologie/phonétique/prosodie, **langues orales ou peu décrites**
- Transcription enrichie (analyse conversationnelle)
- Codages temporels (début et fin)
- Superpositions temporelles (plusieurs locuteurs)
- Indications complémentaires portant sur la situation, …
- Gestes co-verbaux
- Analyse complète de tous les gestes, sons, langagiers ou non

Formats, outils, structure

Corpus sonores, gestuels, d'interactions langagières

Spécificités des corpus audio et visuels
- Les transcriptions ne sont qu'une petite facette du corpus

Audio
- Timing
  - Superposition possible des éléments sonores
- Signal
  - Analyses spectrales
  - Phonétique
  - Phonologie

Vidéo
- Point(s) de vue
  - Une ou plusieurs caméras
- Analyse des éléments visuels
  - Gestes
    - Co-verbaux, langues signées, autres
  - Situation
- Mocap (motion capture)

Corpus sonores, gestuels, d'interactions langagières

Les transcriptions ne sont pas prédéfinies par un système culturel partagé

- l'oral et le gestuel se composent d'énoncés et de mots non segmentés de manière évidente
- les systèmes intègrent de nombreux « accidents » : pauses, reprises, hésitations, etc.
- la représentation de l'oral peut suivre plusieurs conventions dictées par les présupposés théoriques ou le but de la recherche (pas une seule façon de faire)
- la représentation des gestes et des langues signées est en pleine découverte et doit être redéfinie presque pour chaque recherche

- L'absence d'accord sur les transcriptions « impose » le partage des données média
- Une donnée orale/gestuelle sans média n'a pas de saveur
  - Il y manque beaucoup d'information – on ne peut pas garantir l'analyse des données

Création de corpus: exemple du corpus COLAJE

Investissement important humain et matériel sur 10 ans

- Deux ANR, un projet pilote Adonis
- Environ 220 heures de corpus.
- 7 enfants suivis de 11 mois (le plus jeune) à 6 ans 11 mois (le plus âgé)
- 6600 heures de travail de transcription + recueil des données + numérisation/conversion + organisation/contrôle du travail
- 1 400 000 mots, 18 300 mots différents
- Cout estimé minimal de 180 000 euros (1,3 euro pour 10 mots) sans compter le travail complémentaire des chercheurs (encadrement, essais et erreurs)

On retrouverait des chiffres similaires (et plus encore) pour des corpus comme PFC, ESLO, TCOF, CLAPI, CFPP, etc.


MOT: qui c'est qu(i) a acheté ?

Choix de transcription

Format CHAT (CHILDES) pour le partage et pour le codage

Utilisation de CLAN et de PHON (CHILDES)

*MOT: qui c'est qu(i) a acheté ?
*CHI: e@fs papa .
%pho: e papa
*MOT: c'est papa ?
*CHI: oui .
%pho: wi
*MOT: bah non c'est tonton Christophe qui te les a données .
*MOT: c'est tonton Christophe qui t(e) les a données .

*MOT: et il t' a donné quoi d' autre encore ?
*CHI: des carottes et cou(r)gettes .
%pho: de caʁot e cuʒɛt
%act: CHI enlève ses lunettes
*MOT: des carottes et des courgettes ?
%com: MOT prend un air surpris
*CHI: oui .
%pho: wi
%act: CHI remet ses lunettes
*MOT: ah bon ?


MOT: elle est pas belle !

Autre exemple: phonologie

*MOT: elle est pas belle !
*CHI: ‹et celle-ci elle est pas beau› !
%pho: ‹e sɛki el e pa bo›
%mod: ‹e sɛlsi ɛl ɛ pa bo›
*MOT: elle est pas belle on dit .
*CHI: ‹celle-ci elle est pas beau› .
%pho: ‹ʃøsi l e pa bo›
%mod: ‹sɛlsi ɛl ɛ pa bo›
*MOT: Adrien on dit qu'elle est pas belle !
*CHI: ‹est pas beau celle-ci l'est pas beau celle› .
%pho: ‹e pa bo sɛsi le pa bo sɛl›
%mod: ‹ɛ pa bo sɛlsi lɛ pa bo sɛl›
*CHI: ‹celle-ci elle est plutôt› .
%pho: ‹sɛlsi l e pyto›
%mod: ‹sɛlsi ɛl ɛ plyto›
*CHI: yyy .
%pho: gil
%mod: *


MOT: tu voudrais quoi # ça ?

Situation, pointage, actions

Codages complémentaires permettant de préciser la situation, les actions, les analyses comme par exemple celle du pointage

- *OBS: elle est partie chercher les clés ?
- %com: demande de clarification .
- *CHI: ə@fs ta@fs clé .
- %pho: ə ta kle
- %act: CHI quitte subitement le parc et se dirige en direction de ...
- %xpnt: show, avec l'index, en direction d'un jouet .
- *OBS: c'est les clés de quoi ?
- *CHI: yyy po(r)te .
- %pho: əl sapat
- %mod: X pɔʁt
- %int: porte/2/
- %act: CHI s'approche de la petite maison .

Codages spécifiques

Des codages spécifiques de certaines recherches peuvent réalisés directement dans la transcription

- *CHI: ə@fs ta@fs clé .
- %xpnt: show, avec l'index, en direction d'un jouet .
- %xpol: NOUN/clé/ /absent/singular/concrete/inanimate/-/specific/
- /-/-/whole/manipulable/ /-/-/-/-
- *OBS: c'est les clés de quoi ?
- *CHI: yyy po(r)te .
- %int: porte/2/
- %act: CHI s'approche de la petite maison .
- %xpol: NOUN/porte/ /visible(audible)/singular/concrete/inanimate/-/specific/ /-/-/whole/manipulable/ /-/-/-/-

Avantage: le codage peut être réalisé aisément en contrôlant la nature de l’interaction et des données (audio ou vidéo)

Exemple CLAPI – Mondada avec CLAN et CA

/Volumes/M2/CLAPI/lentilles.cha

```
@UTF8
@Begin
@Languages: fra
@Participants:      STE Stella Adult, OPT Opticien Adult
@Options:     CA
@ID:      fra|CLAPI|STE|||||Adult|||
@ID:      fra|CLAPI|OPT|||||Adult|||
@Media: lentilles, audio
*STE:     0 (3.4) &=sonnette •
*STE:     bonjour monsieur↘ •
      (1.1) •
*STE:     bon je voulais bien savoir comment c'est euh ça coûte les: les
      lentilles (0.3) •h dures pa'ce que j'ai des lentilles dures •h •
*OPT:     oui •
*STE:     et je: m' gratte sur une de ça c'est pas si grave mais je voulais
      voulais bien •h savoir ⌈com- ⌉
*OPT:                         ⌊ah pour⌋ savoir le prix (.) ah mais c'est
      madame calumi qui s'en occupe est absente pour ce matin si vous
      voulez r'passer dans l'après midi: elle vous l' dira hein •
*STE:     ah bon ⌈et vous n- vous n' savez⌉ pas
*OPT:           ⌊pa'ce que là c' matin elle est pas là⌋ •
*OPT:     un ordre de prix je pense que::: (0.5) c'est dans l'ordre: c'est des
      d- souples ou des demi souples •
      (0.5) •
*STE:     euh (0.5) qu'est ce que ça veut dire souple: (0.2) ce sont des •
*OPT:     elles sont petites et dures ⌈celles⌉ qu' vous avez
*STE:                               ⌊oui ⌋ •
*OPT:     ⌈oui⌉
*STE:     ⌊oui⌋ oui •
      (0.2) •
09oct20[E|CHAT] 9
```

Outils de transcription et/ou d'analyse

Le travail fondamental a consisté pendant longtemps à écouter sa cassette et prendre son traitement de texte…

- Travail linguistique de qualité (mais à la main) et pas d'analyses sonores ou autres
- Pas de lien direct entre transcription et média

Le travail sur corpus oraux nécessite des outils spécifiques

- Annotations alignées sur le(s) son(s) ou la(les) vidéo(s)
  - Accès direct au son ou à la vidéo
  - Contrôle et réanalyse en direct
  - Analyses des signaux sonores ou visuels
- Codages plus complexes
  - Tableurs
  - Base de données

Exemples d'outils et d'usage – formats textuels

Transcriber – **audio** - logiciel efficace pour la transcription – nécessite d'autres outils pour l'exploitation des corpus (par exemple: TXM)

CLAN – **audio et vidéo** - logiciel efficace pour la transcription – fonctions spécialisés d'exploitation pour ce format et pour le domaine ciblé (acquisition du langage) - une certaine souplesse qui permet d'autres utilisations – export aisé vers d'autres outils en format texte

mais bon non mais sinon oui à peu près euh vingt-cinq ans euh je pense que

voilà

ch_CD2

donc tu as vécu à Saint-Cyr-en-Val

ch_CD2 + GK11

1: pendant ton enfance euh
2: alors euh j'ai vécu à Saint-Cyr-en-Val

GK11

jusqu'à vingt ans à peu près

ch_CD2

d'accord

GK11

mais après j'ai pris un appartement euh
en centre ville d'Orléans et depuis dix ans donc je suis on va dire plus ou moins dans le centre ville

je suis parti euh travailler un petit peu sur Paris je suis revenu sur Orléans euh mais bon
j'ai toujours eu moi mes mes parents sur Orléans donc euh j'ai touj- vraiment toujours un contact à Orléans euh

GK11 + ch_CD2

1: je m'en suis jamais éloigné vraiment quoi voilà
2: d'accord d'accord oui oui oui oui

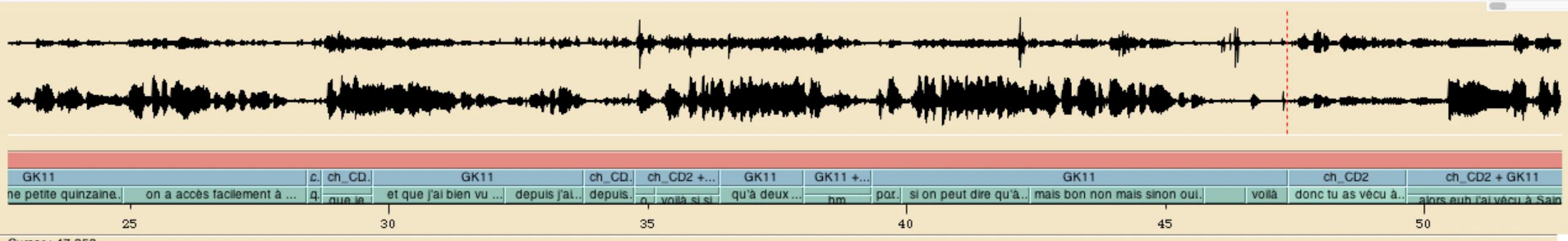
ch_CD2

même pas voilà tu as fait tes études euh

ch_CD2 + GK11

ESLO2_ENT_1010_C

Resolution



Cursor : 47.353

/Users/christopheparisse/Desktop/montpellier/exemples/JULIE-30-3_00_30.cha

@Begin
@Languages: fra
@Participants: CHI Julie Target_Child, MOT Françoise Mother, BRO Ulysse Brother, FAT Nicolas Father
@ID: fra|Paris-Corpus_Julie|CHI|3;00.30|female|||Target_Child|||
@ID: fra|Paris-Corpus_Julie|MOT|||female||Mother|||
@ID: fra|Paris-Corpus_Julie|BRO|4;04.25|male|||Brother|||
@ID: fra|Paris-Corpus_Julie|FAT|||male||Father|||
@Media: JULIE-30-3_00_30-240p, video
@Date: 09-JAN-2010
@Location: CHI's home
@Time Duration: 3745.118
@Situation: CHI et BRO sont filmés à la maison avec MOT et FAT
*MOT: on est le neuf janvier .•
*MOT: on filme Ulysse et Julie à la maison avec papa et maman .•
*MOT: c'est maman qui filme pour le moment .•
*MOT: qu'est+ce+que vous faites ?•
*BRO: moi zə@fs fait d(e) la plasticine et z@fs essaie de faire un petit bonhomme .•
%pho: mwa zə fɛ d la plastisin e z ɛsɛ də fɛʁ œ̃ pəti bɔnɔm
*FAT: +< 0 [=! tousse] .•
*FAT: xxx .•
*BRO: +< veux voir maman .•
%pho: vø vwaʁ mamɑ̃
*MOT: Ulysse j(e) te montrerai après .•
*MOT: d'accord ?•
*MOT: pour le moment je filme .•
*MOT: et toi Julie qu'est+ce+que tu fais chérie ?•
*CHI: de la plasticine .•
%pho: də la plastisin
*MOT: génial !•
*BRO: attends i(l) faut qu(e) zə@fs +//.•
14sep17[E|CHAT] 17

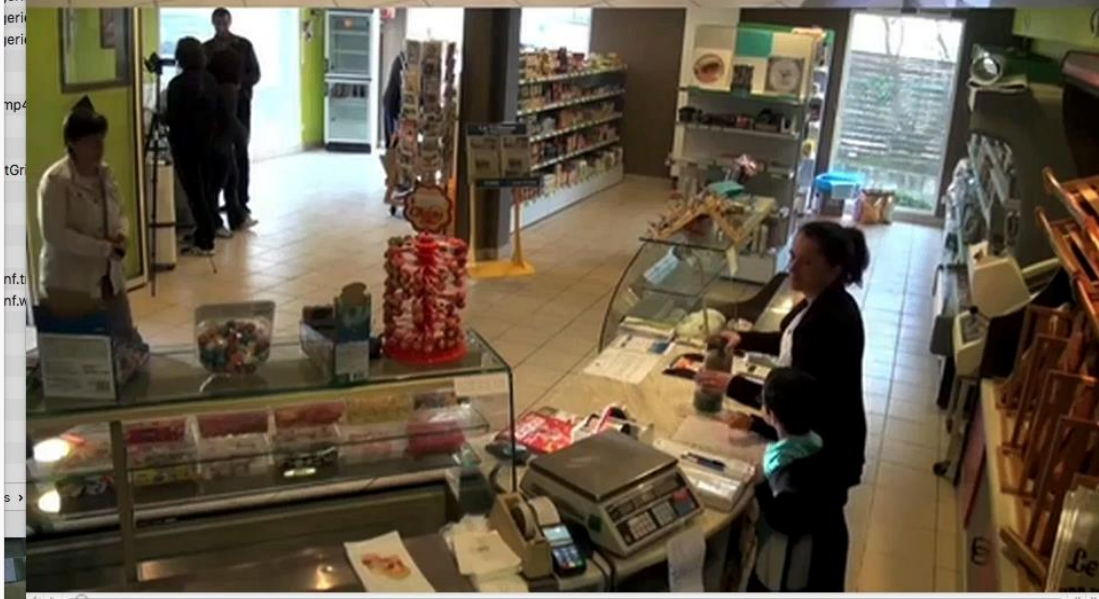
Movie - Sound

1760
Save
v
v
3726650
epeat
0
msec
IE-30-3_00_30-240p.

/Users/christopheparisse/Desktop/montpellier/exemples/CLAPI_commerce_boulangerie_rurale_C01_C20_t...

*: [=!0.9 /instantaneous/N] •
*spk2: j` croyais qu'il allait être plus lourd {(que ça) /instantaneous/LX} •
*: [=!0.4 /instantaneous/N] •
*spk4: °prendre ça° •
*: [=!0.8 /instantaneous/N] •
*spk4: y a du monde [=!/ /previous/PHO] c` matin [=!\ /previous/PHO] •
*spk1: ouais j` suis filmée [=!/ /previous/PHO] _ [=!. /instantaneous/N] alors euh {: /previous/LX} •
*spk4: vous [aussi d'ailleurs_____]_ [=!((rires)) /instantaneous/N] •
*spk1: _____[ah bon [=!\ /previous/PHO] c'est pas très sympa ça [=!/ /previous/PHO]] •
*spk4: c'est la dernière •
*: [=!0.2 /instantaneous/N] •
*spk1: c'est [la dernière [=!/ /previous/PHO]]_ [=!C01 repart vers les rayons /previous/COM] •
*spk2: _____[{(inaud.) /instantaneous/LX} ____] •
*: [=!2.3 /instantaneous/N] •
*spk2: c'est les derniers [=!// /previous/PHO] •
*: [=!3.5 /instantaneous/N] •
*: [=!1.5 /instantaneous/N] _ [=!entrée de C02 /previous/COM] •
@G: report Client 02
*spk1: bon[jou {: /previous/LX} r__] •
*spk5: ___[bonjour] •
*: [=!0.8 /instantaneous/N] •
*spk5: j` voudrais une flûte farinée {: /previous/LX} [=!/ /previous/PHO] et une baguette longue •
*: [=!5.9 /instantaneous/N] _ [=!VE1 prépare la commande /previous/COM] •
*spk1: °alo {: /previous/LX} rs euh {: /previous/LX} {: /previous/LX} °_ [=!VE1 calcule à la calculatrice /previous/C
*spk1: euh {: /previous/LX} DEUX quinze [=!/ /previous/PHO] s'il vous plaît [=!\ /previous/PHO] •
*: [=!8.5 /instantaneous/N] _ [=!VE1 plie la commande; CO2 paie /previous/COM] •
*: [=!0.7 /instantaneous/N] _ [=!VE1 sort la monnaie /previous/COM] •
*spk1: [=!VE1 rend la monnaie /previous/COM] et {(de ;deux) /instantaneous/LX} vingt •
*spk1: [=!0.4 /instantaneous/N] [=!VE1 rend la monnaie /previous/COM] •
*spk1: [merci]_ [=!range l'argent /previous/COM] •
*spk5: [merci]_ [=!range sa monnaie /previous/COM] •
*: [=!1.0 /instantaneous/N] •
*spk1: et bonne [journée {: /previous/LX} _] •
*spk5: _____[bonne jou]rnée •
*spk5: mer[ci au revoir] •
*spk1: ___[_merci au re]voir •
*: [=!3.4 /instantaneous/N] _ [=!sortie de C02 /previous/COM] •
*: [=!20.2 /instantaneous/N] _ [=!VE1 va en fabrication et parle avec MAR /previous/COM] •
*: [=!21.0 /instantaneous/N] _ [=!VE1 revient à la caisse et s'occupe d'une brioche /previous/COM] •
*: [=!0.5 /instantaneous/N] •
*spk4: hop •
*: [=!6.1 /instantaneous/N] _ [=!place la brioche en vitrine /previous/COM] •
*spk1: alors {: /previous/LX} •
*: [=!2.1 /instantaneous/N] _ [=!VE1 pèse les achats /previous/COM] •
*spk3: {(t` es bien calme aujourd'hui) /instantaneous/LX} •
*: [=!1.9 /instantaneous/N] •
*spk4: qu'est-ce qu'il dit [=!/ /previous/PHO] _ [=!en riant /previous/COM] •
*spk4: [=!en riant /begin/PHO] il est {: /previous/LX} [{: /previous/LX} calme [=!/ /previous/PHO]] [=!en riant /end/I
*spk1: _____[i` dit {: /previous/LX} _] •
*spk2: {(inaud.) /instantaneous/LX} _ [=!VE1 pèse les achats /previous/COM] •
*spk1: i` dit t'es bien calme [=!/ /previous/PHO] aujourd'hui [=!\ /previous/PHO] •
*spk1: [[=!((rires)) /instantaneous/N]] •
*spk4: [_____ah] •
*: [=!0.3 /instantaneous/N] •
14sep17[E|CHAT] 89

Movie - Sound



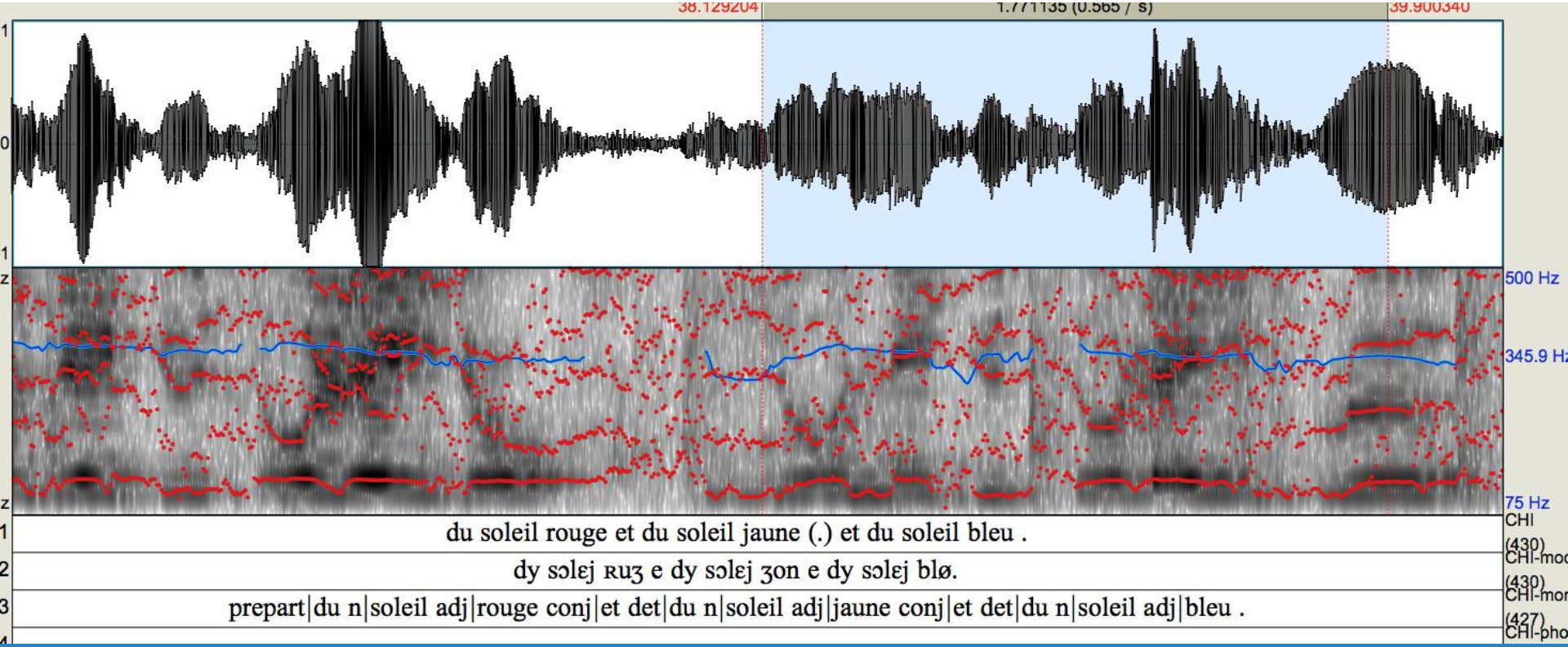
22591 Save
v v
12172 2228520

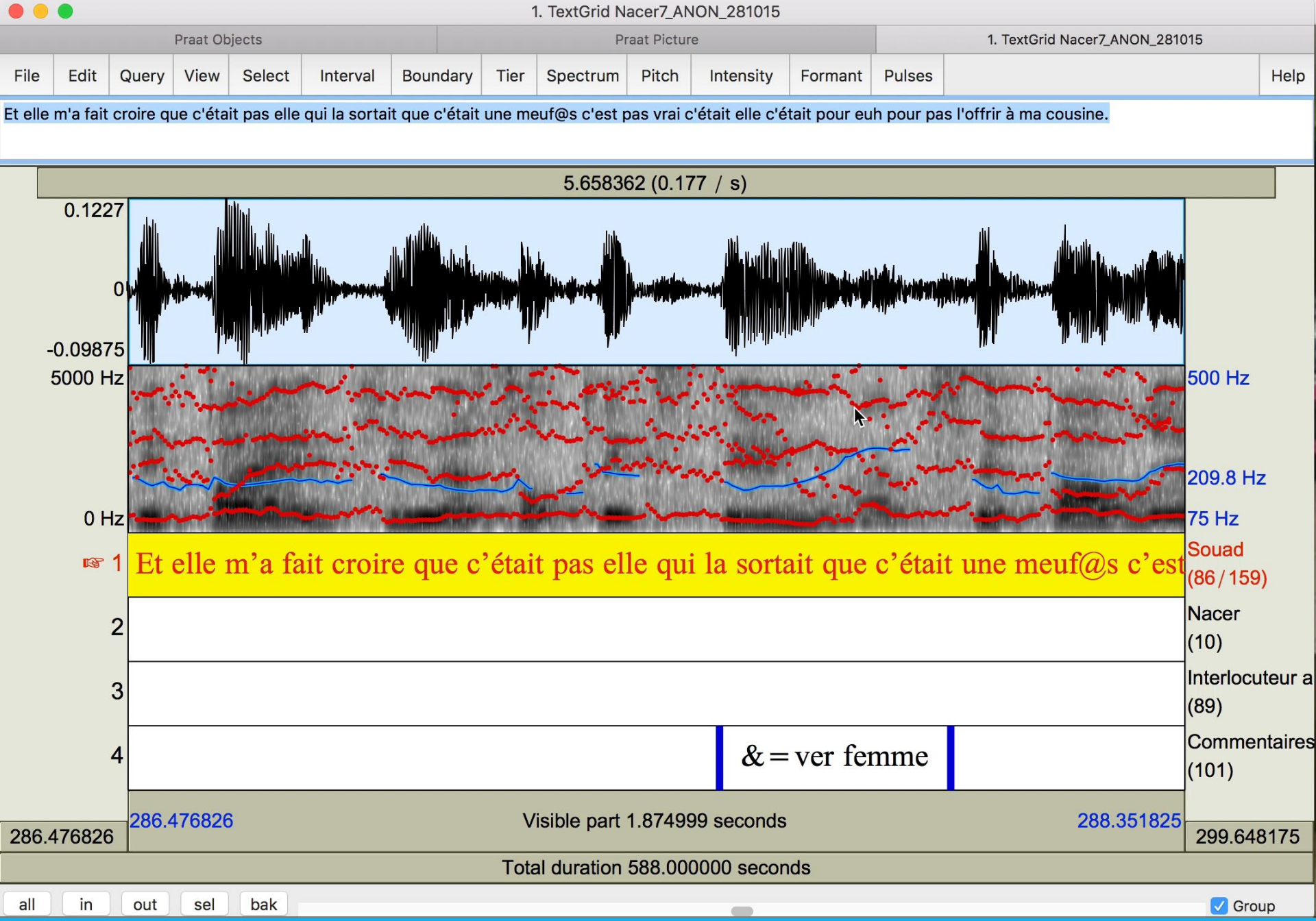
Exemples d'outils et d'usage – phonologie

PHON – **audio et vidéo** – logiciel de codage et d'interrogation de corpus de phonologie (usage texte tout à fait possible – compatible avec CLAN) – voir http://phonbank.talkbank.org/

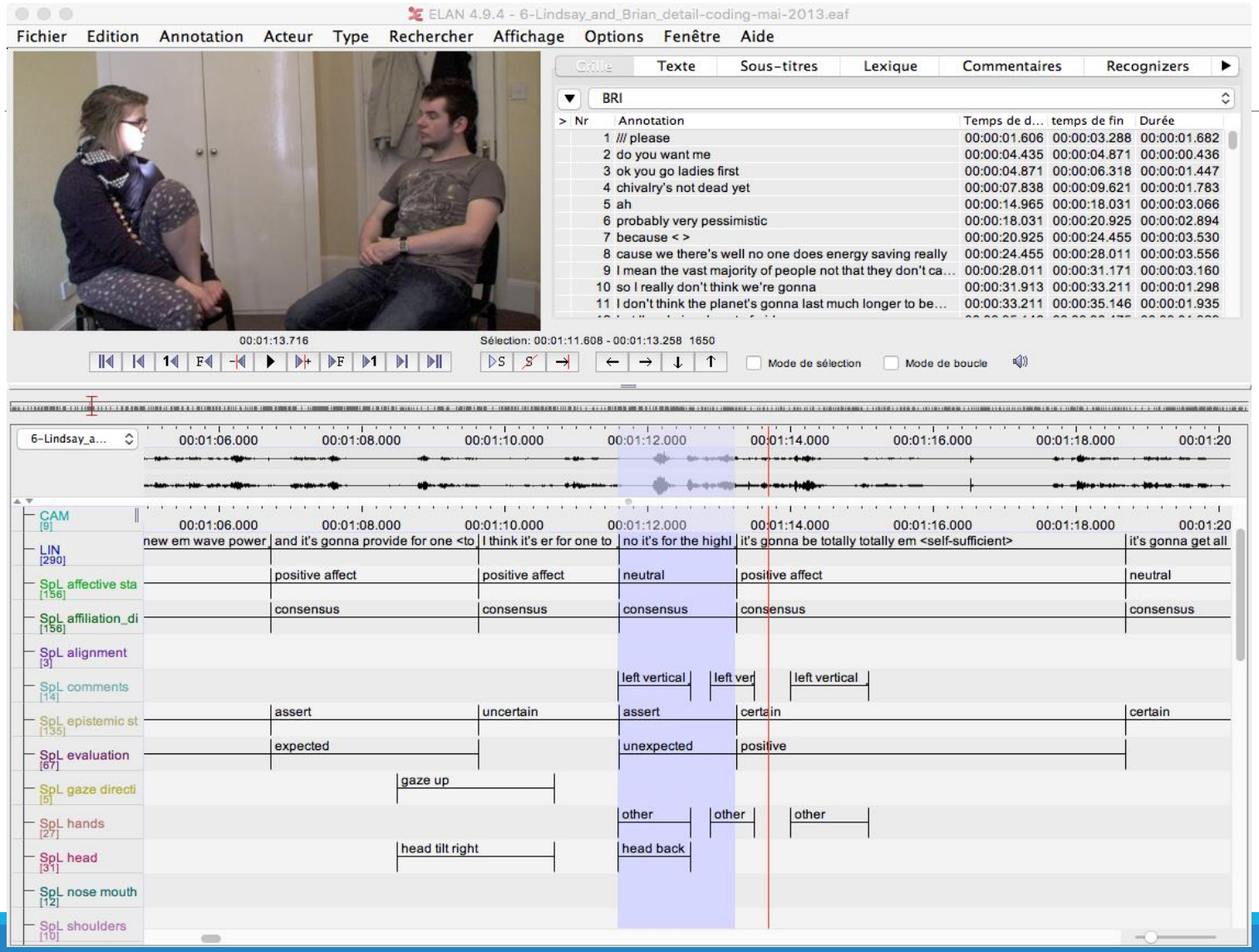
Exemples d'outils et d'usage – formats partition

Praat – **audio** – transcription multi-couche alignée sur le temps – logiciel d'analyse du signal et de phonétique – traitements programmables du signal

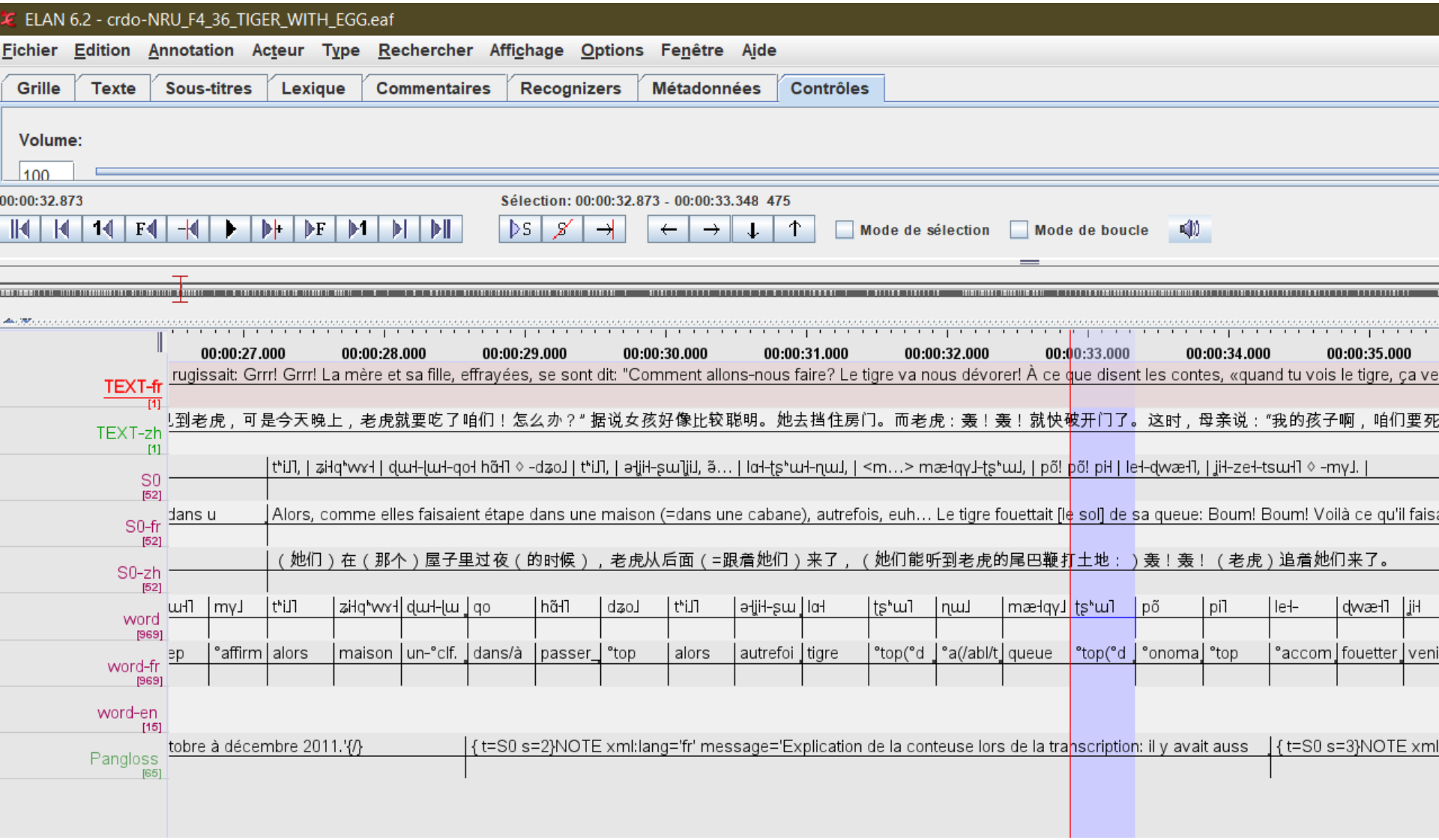



1. TextGrid Nacer7_ANON_281015
Praat Objects
Praat Picture
1. TextGrid Nacer7_ANON_281015
File
Edit
Query
View
Select
Interval
Boundary
Tier
Spectrum
Pitch
Intensity
Formant
Pulses
Help
Et elle m'a fait croire que c'était pas elle qui la sortait que c'était une meuf@s c'est pas vrai c'était elle c'était pour euh pour pas l'offrir à ma cousine.
5.658362 (0.177 / s)
0.1227
0
-0.09875
5000 Hz
0 Hz
500 Hz
209.8 Hz
75 Hz
1
Et elle m'a fait croire que c'était pas elle qui la sortait que c'était une meuf@s c'est
Souad
(86 / 159)
2
Nacer
(10)
3
Interlocuteur a
(89)
4
& = ver femme
Commentaires
(101)
286.476826
Visible part 1.874999 seconds
288.351825
286.476826
299.648175
Total duration 588.000000 seconds
all
in
out
sel
bak
Group

ELAN – **audio et vidéo(s)** – logiciel d'analyse multi-couche et/ou structurée



Donnée langues orales (Pangloss)



Politique scientifique

Coût, partage, diffusion, réutilisation

Créer un corpus de langage oral, quelque soit la complexité des annotations utilisées, est quelque chose de très couteux en temps, en main d'œuvre

Application des principes FAIR

- **Findable, Accessible, Interoperable, Reusable *(Facile à trouver, Accessible, Interopérable, Réutilisable)***

et de la science ouverte

Principe de conservation

Sauvegarder toutes les données originales ou secondaires

- Outils comme ORTOLANG, COCOON, NAKALA
- Collections organisées comme PANGLOSS (langues orales), CHILDES/TALKBANK (child language, other)

Utiliser des formats de longue durée

- Choisir des formats qui seront toujours lisibles dans 10, 20, 50 ans
- En général les formats ouverts, libres, les formats texte (Unicode), XML

Suggestions d'instituts comme le CINES (archives nationales) qui sont à terme susceptibles de réutiliser les corpus dans le futur lointain

Principe de partage

Utiliser des formats connus, décrits, faciles à utiliser pour autrui

Accompagner les dépôts de toutes les informations annexes ou spécifiques + les publications associées

Accompagner de licences spécifiant les réutilisations possibles

Mettre dans des dépôts visibles sur Internet ou sur des catalogues spécialisés avec toutes les métadonnées (données accompagnant les données primaires) permettant d'informer sur l'accès, le partage, la réutilisation

Principe de réutilisation et d'enrichissement

Donner les informations pour la réutilisation et l'enrichissement des corpus

Citer les corpus originaux lors de la réutilisation

Déposer les nouvelles versions (enrichies) en suivant les mêmes protocoles que pour le dépôt original

Base de données CHILDES – depuis 1986

Disponibilité de corpus historiques et de nombreux autres corpus

- 14000 citations du livre de Brown, 1973, – basé sur un corpus disponible

Données présentées dans un format similaire pour tous les corpus

Incitation forte à l'utilisation des données avec citation des articles associés aux corpus

En 2006, 3100 articles utilisant la base de données CHILDES

En 2021, 19700 résultats dans google scholar pour une recherche de « CHILDES »

Exemple corpus CEFC
(projet Orfeo - Debaisieux + 7 laboratoires)

Un corpus Oral de 4 millions de mots constitué à partir de 14 corpus sources de France, Suisse et Belgique.

Téléchargeable sur Ortolang https://hdl.handle.net/11403/cefc-orfeo/v1.5

Interrogeable sur https://orfeo.ortolang.fr/?locale=fr ou sur http://match.grew.fr/

C. Benzitoun, J.-M. Debaisieux, H.-J. Deulofeu (2016). Le projet ORFÉO : un corpus d'études pour le français contemporain. Corpus n°15, p. 91-114

J.-M. Debaisieux & C. Benzitoun (2020). Orféo : un corpus et une plateforme pour l'étude du français contemporain. Langages n°219.

Exemple d'agrégation de corpus (Parisse, Benzitoun, Etienne, Liégeois)

Démontrer la faisabilité de la réutilisation de corpus.

Utilisation du programme de conversion TEICORPO développé pour CORLI.

- Lyon -: CHILDES - Clan
- Eslo -: Cocoon - Transcriber
- Alipe -: Ortolang - Clan
- Colaje -: Ortolang - Clan
- Mpf -: Ortolang - Praat
- Pfc -: Ortolang - Praat
- Tcof -: Ortolang - Transcriber
- Clapi -: Lyon - Transcriber
- Cfpp2000 -: http://cfpp2000.univ-paris3.fr/ - Transcriber

Résultat

Un corpus de 8,9 millions de mots, hors ponctuations.

Analyse syntaxique par TreeTagger avec modèle Perceo

Export vers TXM (export possible vers Iramuteq ou Le Trameur)

Interrogation avec TXM pour répondre à des questions linguistiques demandant de grands corpus

	tous	alipe	cfpp	clapi	colaje	eslo1	eslo2	lyon	mpf	pfc	tcof
Total	10327834	209941	705484	134147	1799932	2225448	1755129	1118081	231227	765646	1382799
Mots	8948165	172949	643253	116063	1457392	2033511	1586838	860243	203054	619493	1255369
Types	78390	6370	18725	8494	18254	27968	29587	11319	10039	18714	27804

Parisse, Benzitoun, Étienne et Liégeois (2021). Agrégation automatisée de corpus de français parlé In : Des corpus numériques à l'analyse linguistique en langues de spécialité [en ligne]. Grenoble : UGA Éditions, 2021. DOI : https://doi.org/10.4000/books.ugaeditions.24220.

Passage de Transciber à TXM

Exemple du corpus ESLO, dont la version actuelle a été transcrite à l'aide de Transcriber

Les métadonnées (informations sur le lieu, la situation, les locuteurs) ont été saisies séparément avec un tableur

Pour exploiter scientifiquement les données du corpus, une conversion avec TEICORPO a permis de créer des fichiers TEI avec des métadonnées riches (issues automatiquement des tableurs)

Les fichiers TEI sont intégrés dans TXM, avec les métadonnées et la possibilité de les écouter dans TXM

Badin, Liégeois, Thiberge, Parisse (2021). Vers un outillage informatique optimisé pour corpus langagiers oraux en vue d'une exploitation textométrique : le cas des interrogatives partielles dans ESLO, Corpus, 22. https://journals.openedition.org/corpus/5421

CORLI : CORpus, Langues et Interactions

- Consortium de l'infrastructure Huma-Num dédié à la linguistique de corpus
- Comité de pilotage de plus d'une vingtaine de personnes et de laboratoires, 180 chercheurs
- Large couverture des différents domaines de la linguistique
- Groupes projet
  - Inter-Explo : Interopérabilité /Pratique et outils d'exploration de corpus
  - Multicom : CMC / LSF et gestualité
  - Corpus multilingues et plurilingues
  - Questions Éthiques & Cadre Juridique
  - Annotation
  - Évaluation des corpus

Buts de CORLI

1. Représenter la communauté
2. Aider la communauté à se structurer (dans son utilisation des outils numériques)
3. Répondre aux besoins de la communauté
4. Diffuser, partager, les outils numériques et les bonnes pratiques
5. Soutenir des actions nécessaires: corpus, outils, guides, formation

CORLI : CORpus, Langues et Interactions

- **Fédérer** les laboratoires travaillant sur corpus pour **recenser** les ressources, les pratiques et les besoins
- **Mutualiser** les ressources et les **diffuse**r dans des modalités et des formats garantissant leur interopérabilité
- **Établir et partager** des bonnes pratiques
- Diffuser des **standards** européens et internationaux : contributions CLARIN , DARIAH et au Consortium TEI
- Établir des **critères d'évaluation des corpus** en tant que production scientifique

Création du centre Knowledge

- CORLI est conscient de l'importance et de l'intérêt des ERIC européens comme CLARIN. Pour cela, CORLI a organisé une journée d'information sur CLARIN le 5 septembre 2017.
- Depuis, des membres de CORLI ont participé à des événements CLARIN européens pour faire connaître le travail du consortium.
- CORLI s'est constitué en centre CLARIN K l'été dernier 2020, et a mis à jour son site web pour en faire un guichet pour les utilisateurs français et européens.

Qu’est-ce que Clarin?

CLARIN (Infrastructure de partage de recherche et de technologies pour le langage) est un consortium d’infrastructure de recherche européen (ERIC).

Son but est de mettre à disposition par un accès unique à tout chercheur européen de données et d’outils concernant le langage (pour toutes les sciences sociales et les humanités).

CLARIN propose donc des données, des outils, des services, des formations, une conférence. L’infrastructure CLARIN est basée sur des centres:

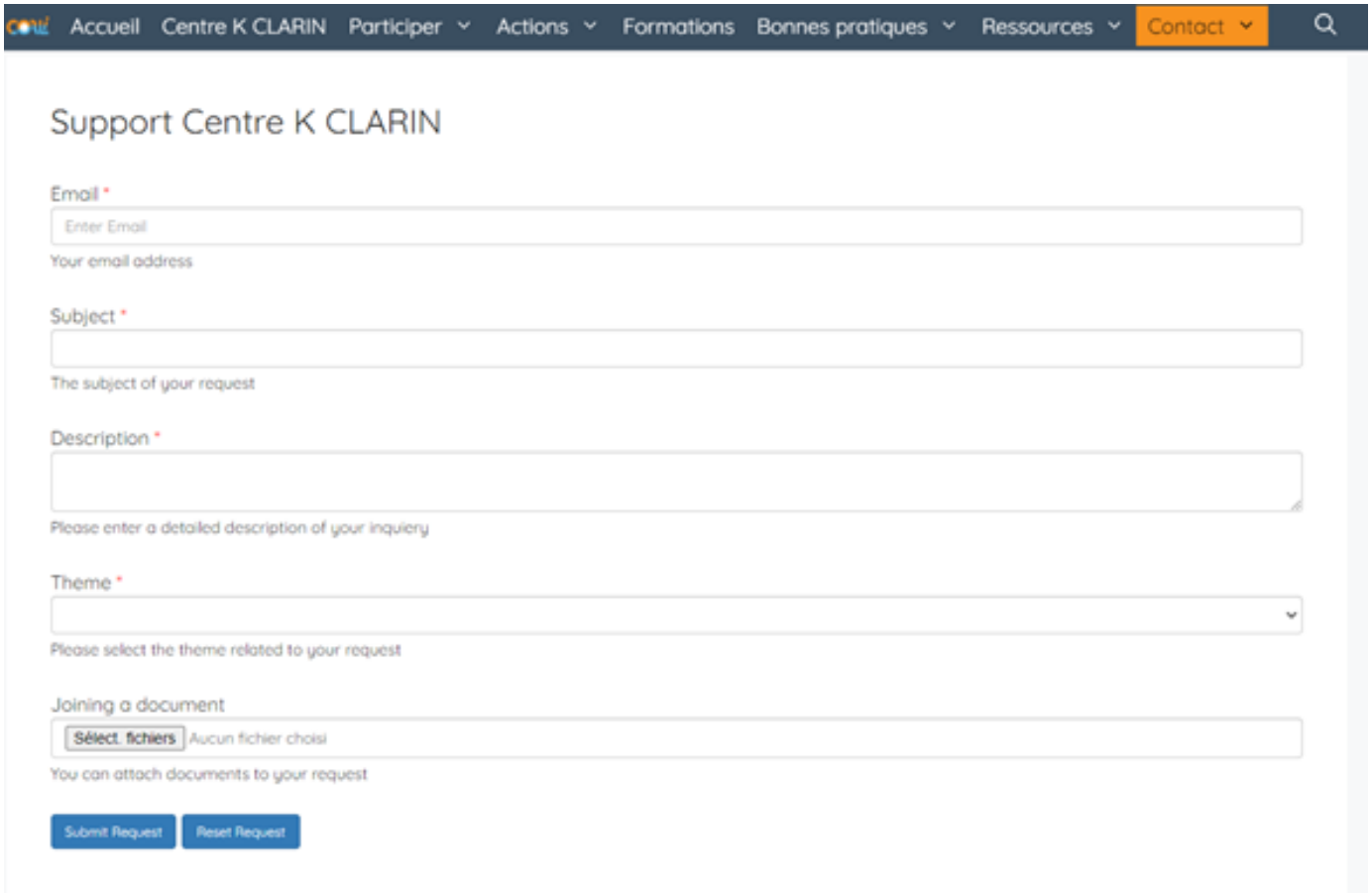
- B ou C pour les données et les outils
- K pour les connaissances, formations, standards, bonnes pratiques, etc.

Corli comme centre K CLARIN

- **Refonte du site web**
  - enrichissement du site
  - mise en place d'une FAQ
  - mise en place d'un système de ticket pour les questions des utilisateurs et des membres
  - traduction du site en anglais

- **Relations avec CLARIN**
  - Clarin FR
  - Participation à la conférence CLARIN 2020 (Soroli et al. 2020)
  - Participation au "TOUR de CLARIN" (présentation par CLARIN de l'implication de la France dans CLARIN)
  - mise en réseau avec les autres centres K

http://corli.huma-num.fr/


Accueil
Centre K CLARIN
Participer
Actions
Formations
Bonnes pratiques
Ressources
Contact
Support Centre K CLARIN
Email *
Enter Email
Your email address
Subject *
The subject of your request
Description *
Please enter a detailed description of your inquiery
Theme *
Please select the theme related to your request
Joining a document
Sélect. fichiers
Aucun fichier choisi
You can attach documents to your request
Submit Request
Reset Request

Perspectives et ambitions pour le centre K et pour CORLI

Lier et partager l'information : anticiper et répondre aux besoins des linguistes, ingénieurs, enseignants, ...:

1. se tenir au courant des nouveautés et des progrès techniques
2. éviter de faire les mêmes choses plusieurs fois
3. diffuser et favoriser l'innovation
4. faciliter les synergies
5. améliorer la reproductibilité des recherches
6. aider à une recherche de haute qualité
7. construire collectivement connaissances et techniques

Perspectives pour CORLI

- **Redéfinir un projet pour 2022-2025 auprès d'Huma-Num**
- Un fonctionnement réseau, un lien avec les linguistes et les producteurs/utilisateurs de corpus (centre K)
- Des groupes de réflexion, de l'information et de la formation autour des outils et des méthodes FAIR
- Maintenir, améliorer les services auprès des linguistes, faire le lien et tirer parti des possibilités offertes par CLARIN
- Développer de nouveaux outils ou techniques répondant à des besoins, en particulier dans la réutilisation des corpus
  - Annotation
  - Citation