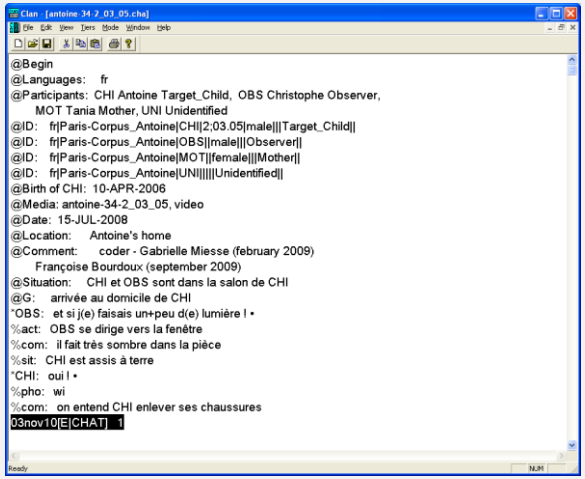
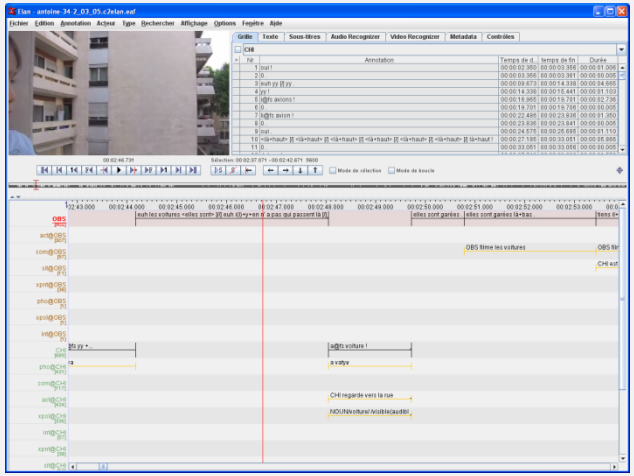
Utiliser, traiter et interroger des corpus existants





Quelques pistes…

- Conversions, échanges, réutilisation d'anciens corpus
- Enrichissement des codes
  - Analyse syntaxique
  - Codage fin et analyse des codes
- Traitements de corpus
  - Fréquences
  - Concordances
  - Cooccurrences
  - Visualisation, navigation
  - Statistiques

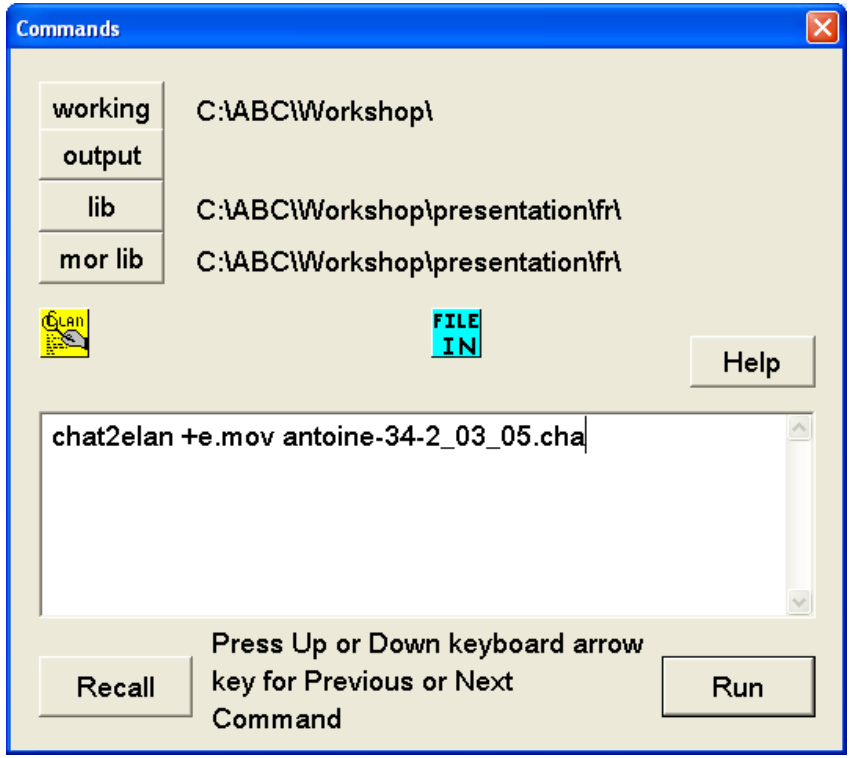
Conversion, échange, extraction

- La plupart des outils disposent de moyens de conversions entre formats
  - CLAN peut importer depuis et exporter vers:
    - PRAAT : **chat2praat** et **praat2chat**
    - ELAN : **chat2elan** et **elan2chat**
    - ANVIL : **anvil2chat** et **chat2anvil**
    - PHON (à travers Chat-Xml = Chatter)
    - Exportation vers EXMARALDA : **chat2xmar**
    - Importation depuis LIPP : **lipp2chat**

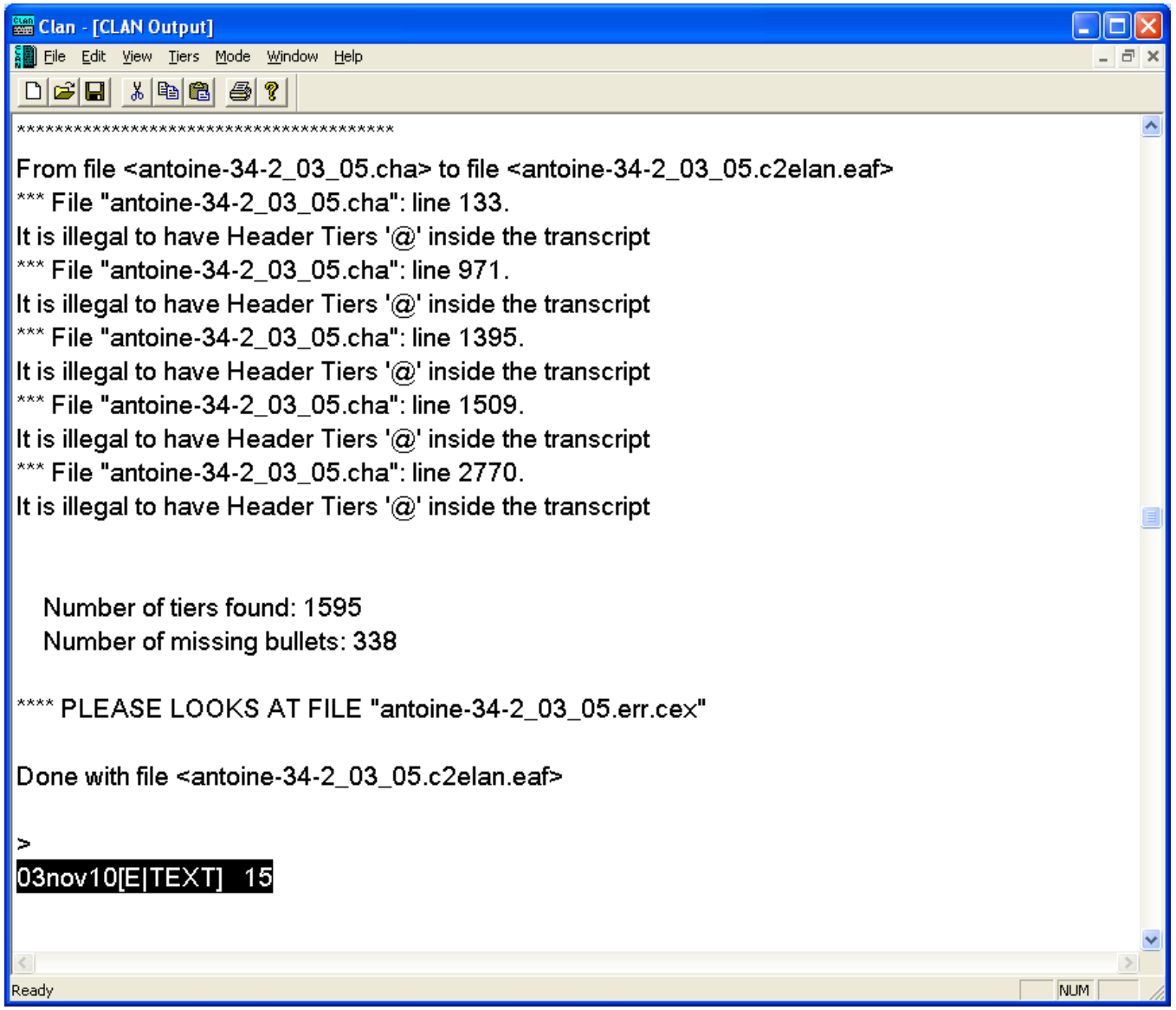
- Exemples dans : **http://www.modyco.fr/corpus/colaje.php**

Conversion de CLAN vers ELAN

- Utiliser la commande avec l'information de l'extension du fichier média

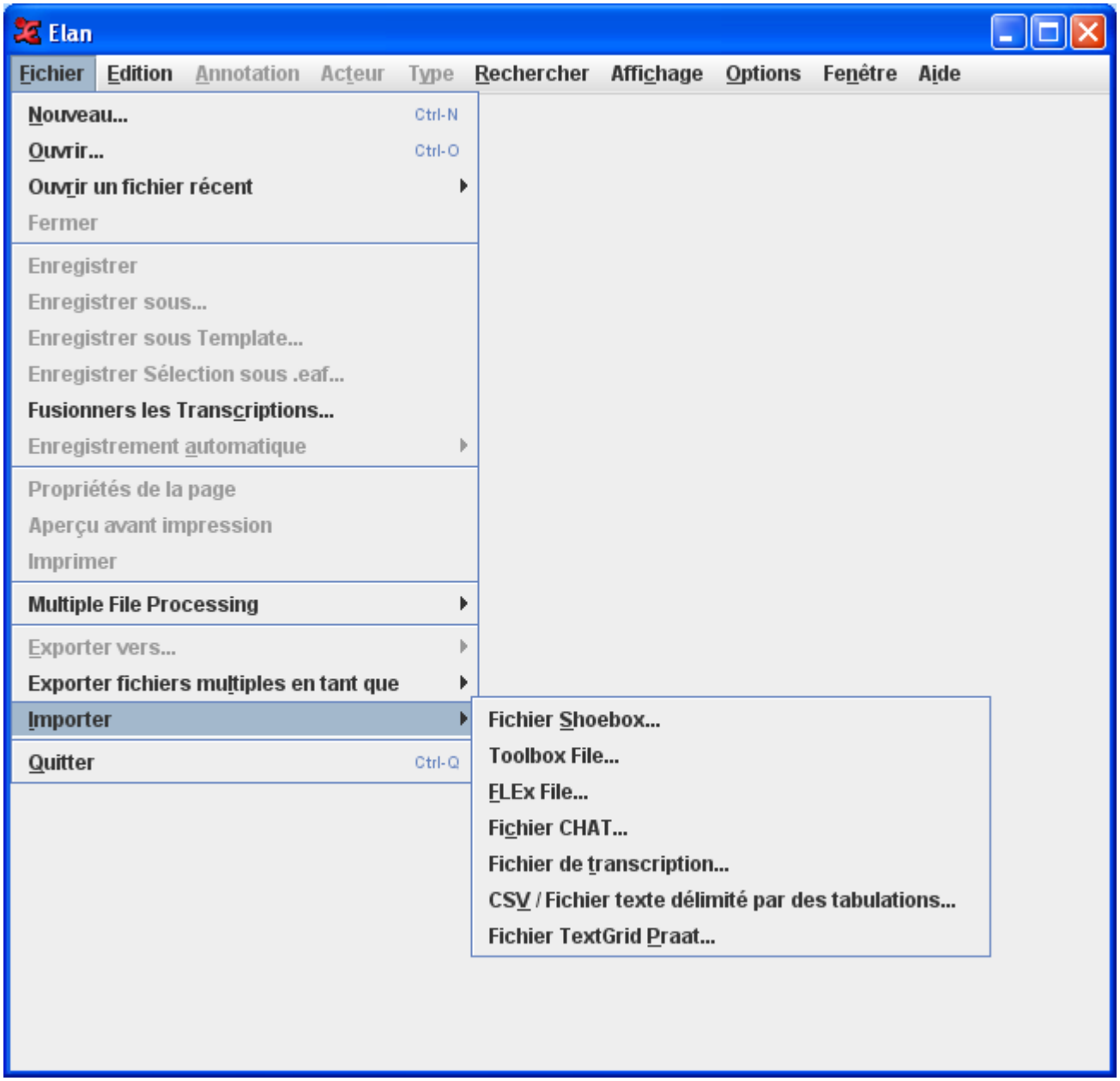


- Messages d'erreurs ne sont pas gênants
- Limites dans les formats et les conversions
- Pertes de données possible (garder l'original!)



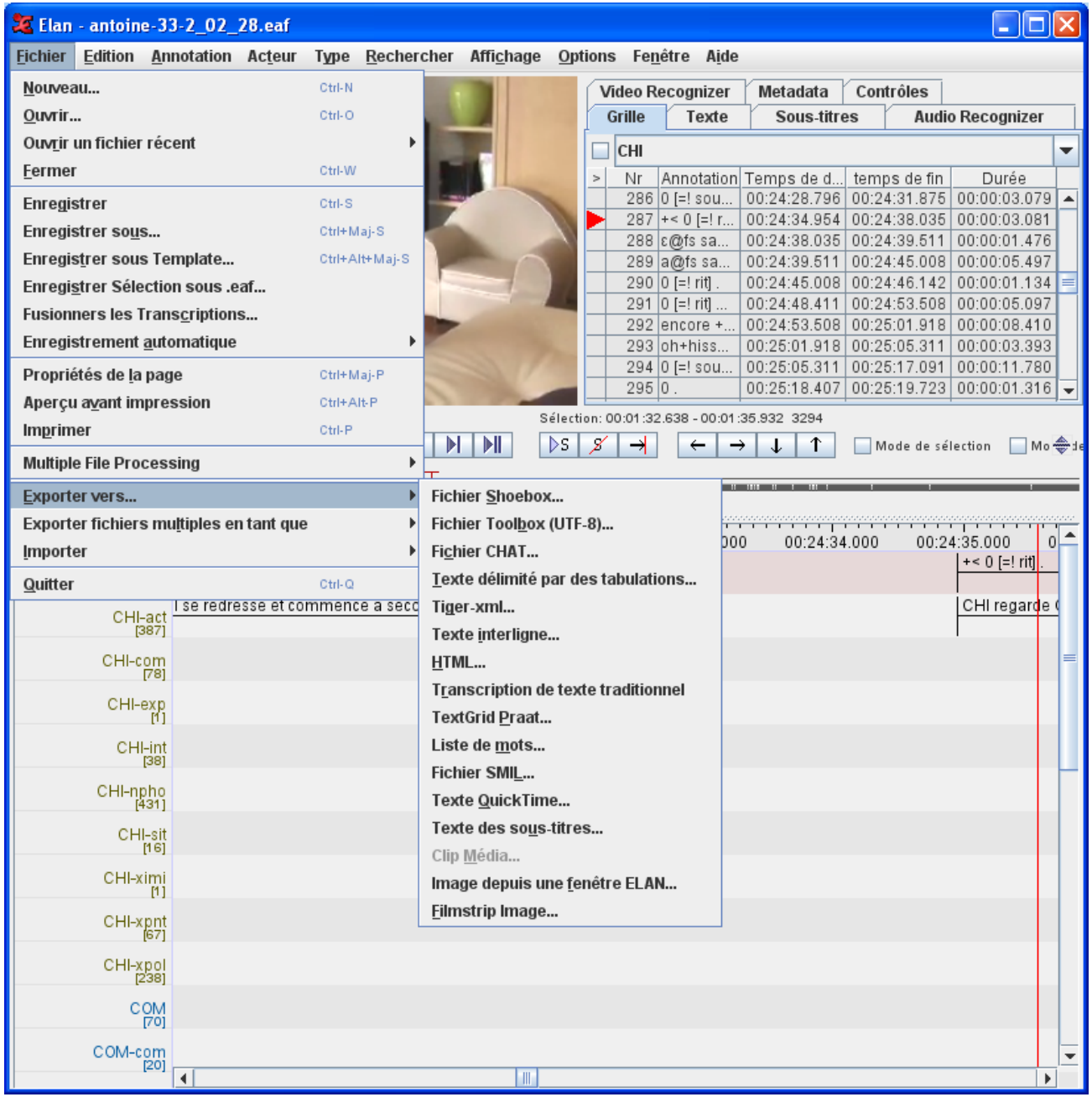
En utilisant ELAN

Importer vers ELAN



En utilisant ELAN

Exporter depuis ELAN



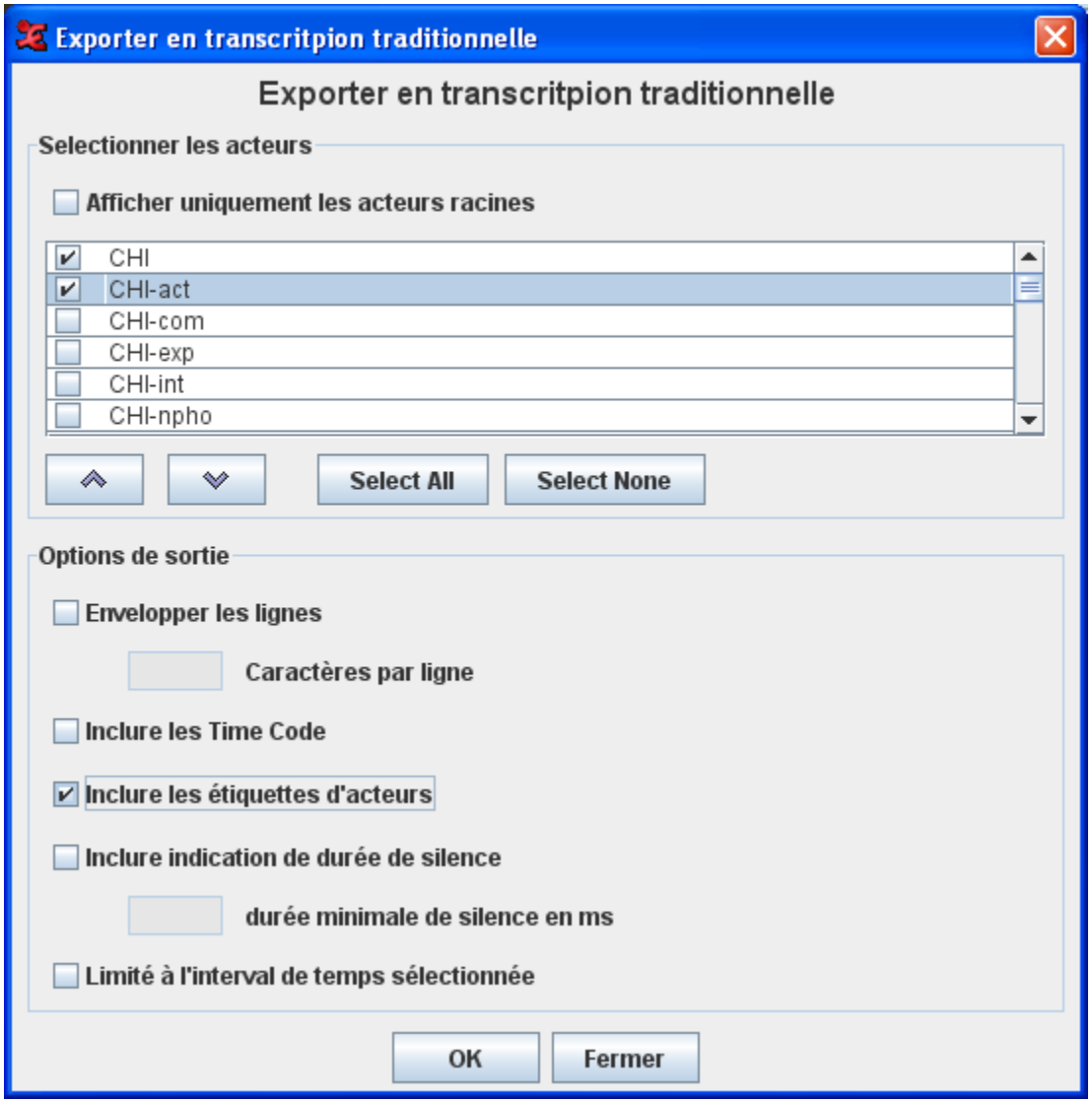
Extraction depuis les corpus

- Récupérer une partie des données
  - Ligne principale
  - Un seul locuteur
  - Un codage (tier) spécialisé
    - Le codage morphosyntaxique
    - Le codage phonologique
- Pour les utiliser avec d'autres outils
  - Tableurs
  - Logiciels de textométrie
  - R, outils de programmation (awk, perl)
  - Création de nouveaux formats

Extraction de parties

- Dans ELAN: Interface graphique détaillée
  - Souple mais non réutilisable
- Depuis CLAN: Interface par ligne de commande
  - Moins facile à utiliser mais plus facile d'automatiser ou de stocker les commandes de traitements

ELAN


Exporter en transcritpion traditionnelle
Exporter en transcritpion traditionnelle
Selectionner les acteurs
Afficher uniquement les acteurs racines
CHI
CHI-act
CHI-com
CHI-exp
CHI-int
CHI-npho
Select All
Select None
Options de sortie
Envelopper les lignes
Caractères par ligne
Inclure les Time Code
Inclure les étiquettes d'acteurs
Inclure indication de durée de silence
durée minimale de silence en ms
Limité à l'interval de temps sélectionnée
OK
Fermer

CLAN

- KWAL permet d'extraire des éléments
  - Extraction d'un champ (locuteur ou autre)
    - Suppression des retours à la ligne

  longtier mar35a.cha
    - Nettoyage des codes

  flo +d mar35a.longtr.cex ; optionel
    - Extraction des données

  kwal +tCHI **+d** +f mar35a.longtr.flo.cex

  kwal **+d** +f mar35a.longtr.flo.cex

  kwal +t%mor +tCHI **+d** +f mar35a.longtr.flo.cex

  kwal +t%mor –t* **+d** +f mar35a.longtr.flo.cex

  voir extrait_kwal.cl

Récupérer d'anciens corpus

- De nombreux corpus existent en format électronique de type Word dans des formes souvent proches des formats utilisés dans les outils
  - Il est souvent possible d'utiliser ces corpus et de les mettre dans un format utilisable par les outils d'alignement
  - Un corpus déjà transcrit peut être facilement aligné en utilisant CLAN
  - CLAN en tant qu'outil ayant une présentation « textuelle » est souvent une bonne passerelle pour aller après vers d'autres outils

De nombreuses commandes outils

- CHSTRING Changes words and characters in CHAT files.
- COMBTIER Combines extra commentary lines.
- CP2UTF Converts ASCII files to Unicode files.
- DATACLEAN Updates the format of older CHAT files.
- DELIM Inserts periods when final delimiters are missing.
- FIXBULLETS Repairs bullets and reformats old style bullets
- FIXIT Breaks up tiers with multiple utterances.
- INDENT Aligns the overlap marks in CA files.
- LONGTIER Removes carriage returns to make long lines.
- **LOWCASE Converts uppercase to lowercase throughout a file.**
- QUOTES Moves quoted material to its own tier.
- REPEAT Inserts postcodes to mark repeated utterances.
- RETRACE Inserts retrace markers.
- SILENCE Converts utterances with LASTNAME to silence
- **TEXTIN Converts straight text to CHAT format.**
- TIERORDER Rearranges dependent tiers into a consistent order.
- UNIQ Sorts lexicon files and removes duplicates.

Enrichir les corpus

- Créer un corpus est souvent une opération manuelle
  - Permet d'intégrer directement l'évaluation de l'utilisateur
  - Seule technique possible actuellement pour la sémantique, la pragmatique
- Des techniques automatiques sont envisageables et intéressantes
  - pour diminuer le travail nécessaire sans interdire la vérification humaine
  - pour créer de très grands corpus pour lesquels la statistiques et la quantité compenseront les erreurs locales

Futur proche

- Reconnaissance de la parole
  - Étiquetage automatique pour constituer des gros corpus d'apprentissage
  - Difficile à mettre en œuvre sur le très jeune enfant, nécessite une adaptation au locuteur (seul blocage probablement le marché – coût, rentabilité, logiciels pas gratuits)
- Identification des silences et des tours de parole, sélection de locuteurs
  - Existe déjà sur certains produits commerciaux (Lena utilisé pour recueillir 1400 jours d'enregistrements et 3 millions d'énoncés dans une étude de Oller & al. sur l'autisme)

Présent

- Analyse syntaxique
  - Il existe de nombreux systèmes et différents niveaux d'analyses
  - Il existe des logiciels commerciaux (pour les correcteurs d'orthographe) et des produits universitaires
    - CLAN
      - MOR+POST : partie du discours, morphosyntaxe
        - » Disponible en français
      - GRASP : syntaxe, fonctions syntaxiques, structure d'énoncé
    - TreeTagger (utilisé dans Le Trameur)
    - Cordial (commercial)

Utilisation de l'analyse syntaxique sous CLAN (Mor + Post)

- Mor : analyse **mor**phologique → fournit toutes les informations lexicales pour un mot hors contexte
- Post : **P**art **O**f **S**peech **T**agger → à partir du résultat de Mor réduit l'ensemble des propositions syntaxiques à une seule proposition (calcul effectué en fonction du contexte → difficile pour les mots isolés !)

*LOC: tu veux faire à manger ?

*LOC: bah oui xx les autres ils dorment encore .

mor mp.cha → mp.mor.cex

*LOC: tu veux faire à manger ?

%mor: pro:subj|tu&2S^v|taire-PP-_MASC v:mdllex|vouloir&IMP&2SV^v:mdl|vouloir&IMP&2SV^v:mdllex|vouloir&PRES&2SV^v:mdl|vouloir&PRES&2SV^v:mdllex|vouloir&PRES&1SV^v:mdl|vouloir&PRES&1SV v:mdllex|faire-INF^v:mdl|faire-INF prep|à^prep:art|à^adv|à n|manger&_MASC^v|manger-INF ?

*LOC: bah oui xx les autres ils dorment encore .

%mor: co|bah n|oui&_MASC^adv:yn|oui unk|xx pro:obj|les&PL^det|les&PL pro|autres&_PL^det:gen|autres&_PL^adj|autre&_PL-MASC^adj|autre&_PL-MASC^adj|autre&_SING-_PL pro:subj|ils&MASC&_3P v|dormir-SUBJV+PRES-3PV^v|dormir-PRES-3PV adv|encore .

post mp.mor.cex → mp.mor.pst.cex

*LOC: tu veux faire à manger ?

%mor: pro:subj|tu v:mdl|vouloir v:mdllex|faire-INF prep|à v|manger-INF ?

*LOC: bah oui xx les autres ils dorment encore .

%mor: co|bah adv:yn|oui unk|xx det|les pro|autres pro:subj|ils v|dormir adv|encore

Installer un dictionnaire

- **fr.zip** (version actuelle) ou fra.zip (futures versions)
- Création d'un répertoire
  - fr (par exemple) contenant:
    - Un ensemble de fichiers décrivant la grammaire (morphologie et syntaxe) et les codes @ (a@l, boum@o)
    - Un sous-répertoire pour le lexique
      - Ces fichiers lexique peuvent être édités par CLAN
      - Il est possible de se créer son propre dictionnaire personnel complémentaire

Utilisation

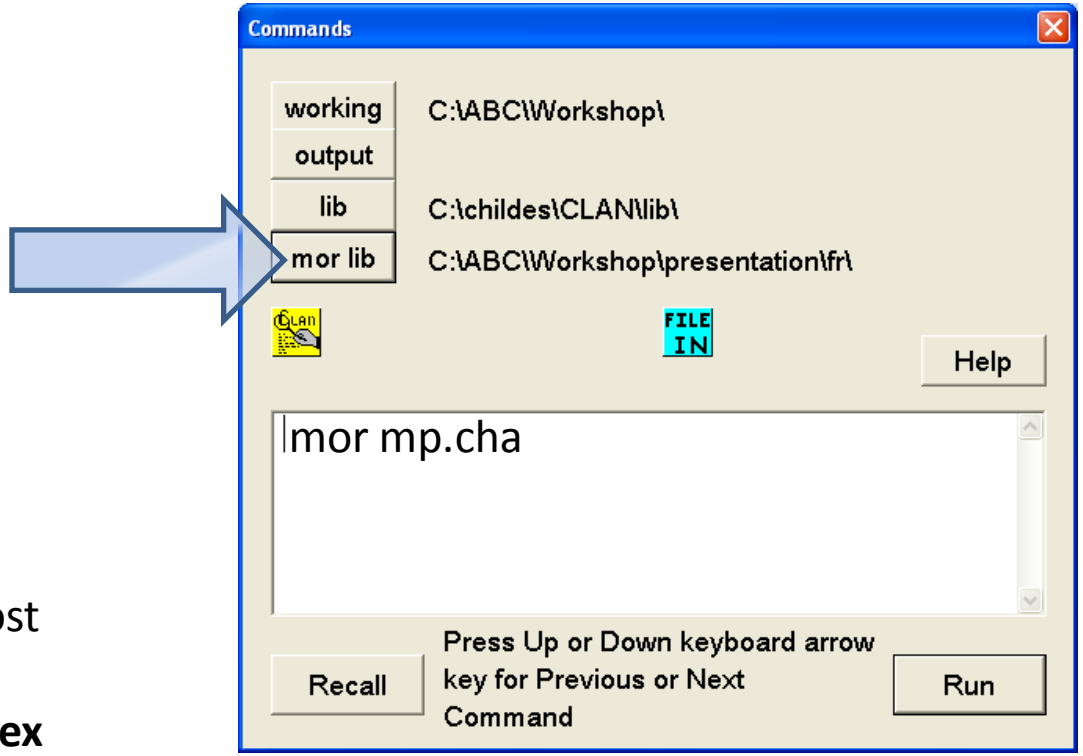
- Ce répertoire doit être « pointé » par la case « Mor lib » de l'interface CLAN

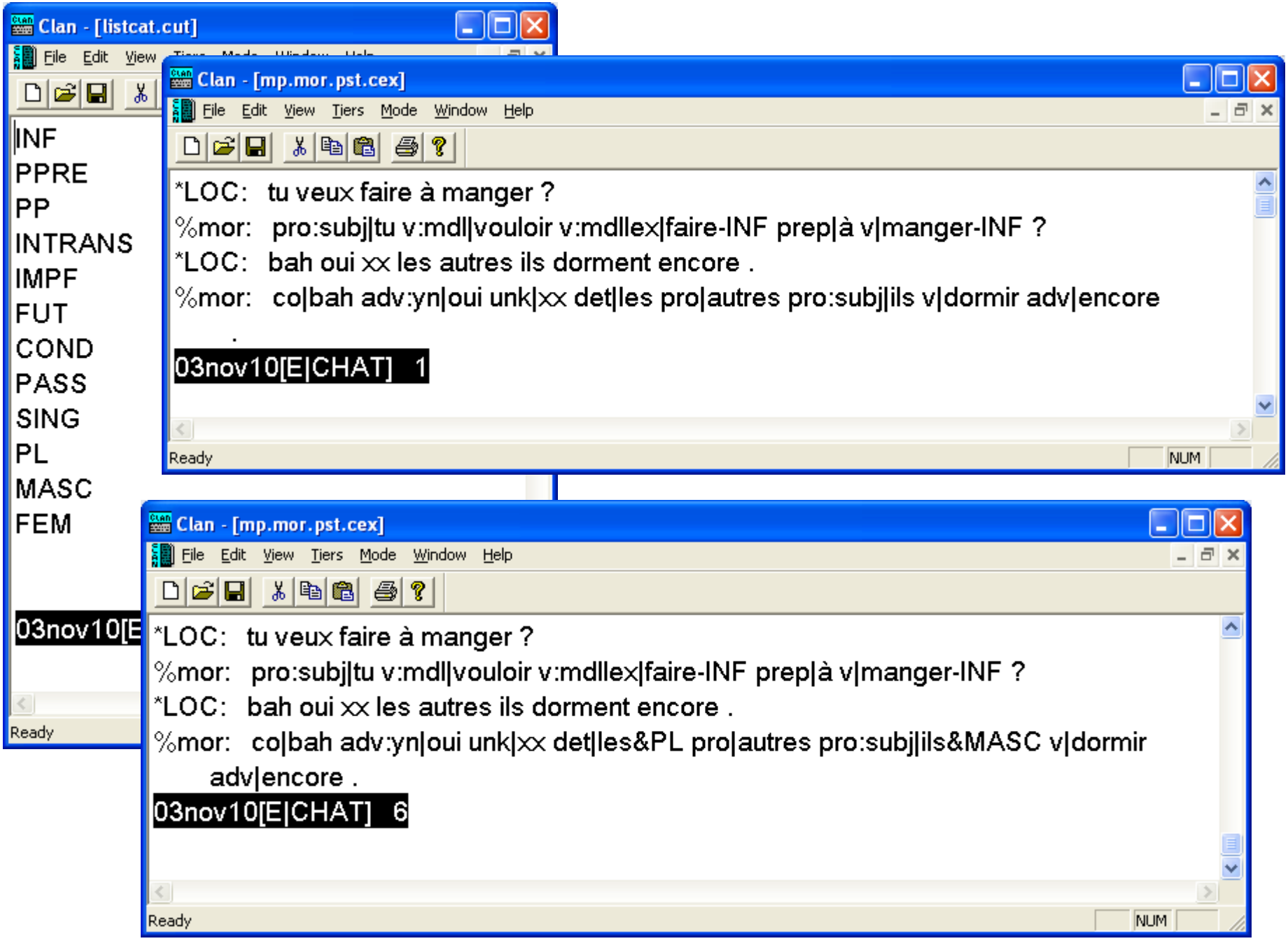
Exécuter en succession

mor mp.cha
post mp.mor.cex

Pour varier la nature des catégories conservées par Post

post +clistecat.cut mp.mor.cex




Clan - [listcat.cut]
File Edit View
INF
PPRE
PP
INTRANS
IMPF
FUT
COND
PASS
SING
PL
MASC
FEM
03nov10[E
Ready
Clan - [mp.mor.pst.cex]
File Edit View Tiers Mode Window Help
*LOC: tu veux faire à manger ?
%mor: pro:subj|tu v:mdl|vouloir v:mdllex|faire-INF prep|à v|manger-INF ?
*LOC: bah oui xx les autres ils dorment encore .
%mor: co|bah adv:yn|oui unk|xx det|les pro|autres pro:subj|ils v|dormir adv|encore
.
03nov10[E|CHAT] 1
Ready
NUM
Clan - [mp.mor.pst.cex]
File Edit View Tiers Mode Window Help
*LOC: tu veux faire à manger ?
%mor: pro:subj|tu v:mdl|vouloir v:mdllex|faire-INF prep|à v|manger-INF ?
*LOC: bah oui xx les autres ils dorment encore .
%mor: co|bah adv:yn|oui unk|xx det|les&PL pro|autres pro:subj|ils&MASC v|dormir
adv|encore .
03nov10[E|CHAT] 6
Ready
NUM

Codage fin

- Les lignes dépendantes peuvent recevoir n'importe quel type de codes
  - phonologie, syntaxe, interaction, gestes
  - tout codage spécifique

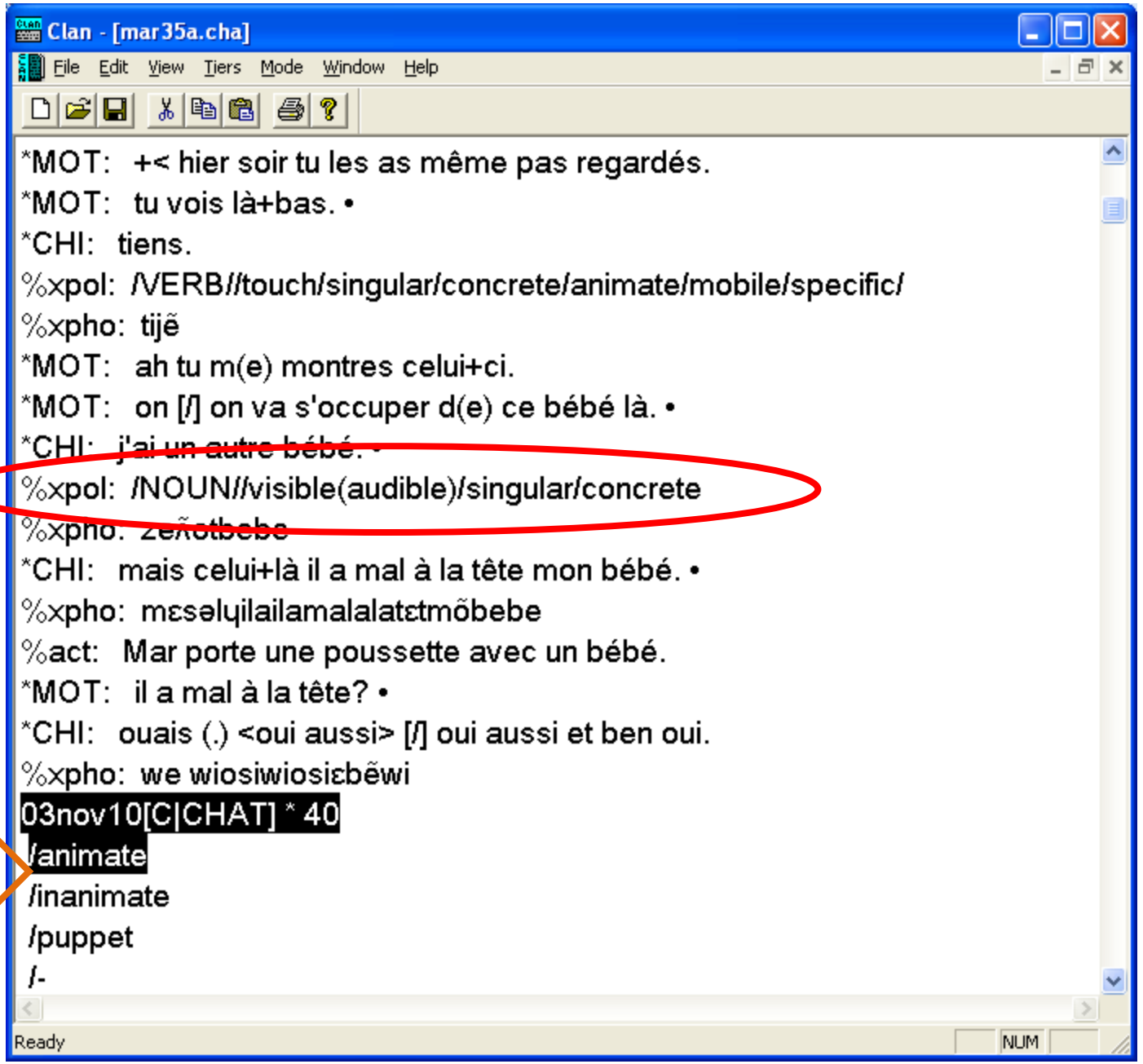
*CHI: musique .
%pho: mœzik
%xfil: musique myzik N bi c m9zik s bi c e-
%xpol: NOUN//visible(audible)/singular/abstract/inanimate/-/-/
%xpol: NOUN/musique/
/visible(audible)/singular/abstract/inanimate/-/-/
/experiencer/-/whole/non-manipulable/ /-/-/-/-/

Outil de codage de CLAN

- Mode coder:
  - Préparation du fichier de description des codes
  - Passage dans le mode coder: Esc-e (choix du fichier de codage)
  - Recherche de l'énoncé à coder: Ctrl+t
  - Choix d'un code: souris ou flèches
  - Validation d'un code: Entrée
    - Saut dans les codes: Esc-s
  - Arrêt du codage: Esc-c + clic en dehors de la zone

Esc e
Ctrl+t
Entrée
Esc s

Esc c
Esc c



Extraire des éléments, des codes, et utiliser un tableur

- L'accessibilité des éléments codés dans CLAN est limitée par le logiciel:
  - Nombre limité de lignes
  - Affichage séquentiel
  - Pas de tri, de statistiques

- Il est possible d'utiliser un tableau pour compléter le travail et de faire des aller-retour entre CLAN et un tableur

Commande KWAL
(ou extraction de champs ELAN)

- KWAL permet d'extraire des éléments
  - Extraction d'une ligne pour utilisation dans Excel
    - Suppression des retours à la ligne

longtier mar35a.cha

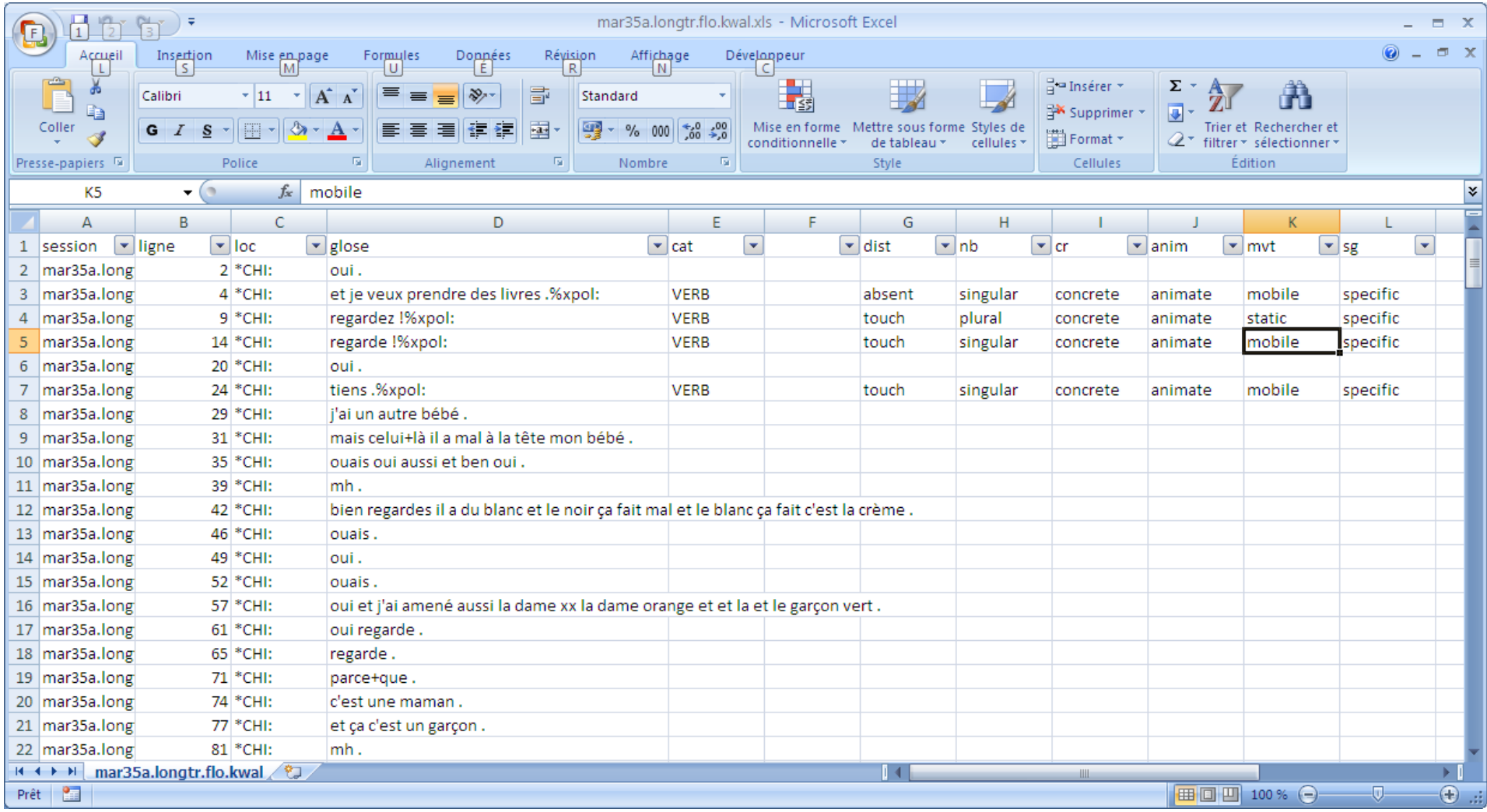
- Nettoyage des codes

flo +d mar35a.longtr.cex

- Extraction de la ligne

kwal +t%xpol +tCHI **+d4** +f mar35a.longtr.flo.cex

voir extrait_mar35a.cl



	A	B	C	D	E	F	G	H	I	J	K	L
1	session	ligne	loc	glose	cat		dist	nb	cr	anim	mvt	sg
2	mar35a.long	2	*CHI:	oui .								
3	mar35a.long	4	*CHI:	et je veux prendre des livres .%xpol:	VERB		absent	singular	concrete	animate	mobile	specific
4	mar35a.long	9	*CHI:	regardez !%xpol:	VERB		touch	plural	concrete	animate	static	specific
5	mar35a.long	14	*CHI:	regarde !%xpol:	VERB		touch	singular	concrete	animate	mobile	specific
6	mar35a.long	20	*CHI:	oui .								
7	mar35a.long	24	*CHI:	tiens .%xpol:	VERB		touch	singular	concrete	animate	mobile	specific
8	mar35a.long	29	*CHI:	j'ai un autre bébé .								
9	mar35a.long	31	*CHI:	mais celui+là il a mal à la tête mon bébé .								
10	mar35a.long	35	*CHI:	ouais oui aussi et ben oui .								
11	mar35a.long	39	*CHI:	mh .								
12	mar35a.long	42	*CHI:	bien regardes il a du blanc et le noir ça fait mal et le blanc ça fait c'est la crème .								
13	mar35a.long	46	*CHI:	ouais .								
14	mar35a.long	49	*CHI:	oui .								
15	mar35a.long	52	*CHI:	ouais .								
16	mar35a.long	57	*CHI:	oui et j'ai amené aussi la dame xx la dame orange et et la et le garçon vert .								
17	mar35a.long	61	*CHI:	oui regarde .								
18	mar35a.long	65	*CHI:	regarde .								
19	mar35a.long	71	*CHI:	parce+que .								
20	mar35a.long	74	*CHI:	c'est une maman .								
21	mar35a.long	77	*CHI:	et ça c'est un garçon .								
22	mar35a.long	81	*CHI:	mh .								

Ajouter d'autres lignes pour les cas d'entrées multiples:

kwal +t%xpol1 –t* +tCHI +d4 +f mar35a.longtr.flo.cex

Autres usages

- Créer des commandes d'extraction personnalisées (R, Perl)
- Basculer les tableaux de type Calculateurs (Excel) en format CSV ou Tabulation pour les utiliser dans un logiciel de statistique
- Utiliser Excel/OpenOffice pour réaliser des statistiques simples, des tableaux dynamiques, des graphiques

Utilisation des corpus

Fréquences, concordances, cooccurrences

- CLAN possède des outils pour obtenir
  - des fréquences : FREQ
  - des concordances et des cooccurrences : COMBO

- Les usages plus complexes ou plus spécifiques imposent de passer à des logiciels de textométrie ou des logiciels de programmation (R, Perl)

Les fréquences avec CLAN

- Commande FREQ
  - Usage simple: **freq** *nom_de_fichier*
- Usages plus complexes
  - Regrouper le lexique de tout un ensemble de fichiers, pour un certain locuteur
    - freq +u +tCHI antoine*.cha
  - Récupérer le résultat
    - freq +u +tCHI antoine*.cha > antoine.txt
  - Liste des nombres de mots
    - freq +d3 +tCHI –sxx* –syy* antoine*.cha
  - Tableau de tous les mots pour tous les fichiers
    - freq +d2 +tCHI –sxx* –syy* antoine*.cha
    - freq +d2 +tCHI +a* antoine*.cha

Les fréquences sur la ligne syntaxique

freq -t* +tCHI +t%mor +u *.mor.pst.cex

1 v|toucher-INF
2 v|toucher-PP
1 v|touiller
2 v|tourner
2 v|tousser
1 v|tousser-PP
3 v|travailler

freq -t* +tCHI +s"@|-*,o-%" +t%mor +u *.mor.pst.cex

792 v
102 v:aux
105 v:exist
101 v:mdl
54 v:mdllex

freq -t* +tCHI +s"@r-*,o-%" +t%mor +u *.mor.pst.cex

6 touche
9 toucher
1 touiller
1 toujours
3 tour

freq -t* +tCHI +s"@&-*,r-%" +t%mor +u *.mor.pst.cex

27 v&INF
4 v&PP
4 v:mdllex&INF

Les concordances avec CLAN

- Commande COMBO
  - Permet de chercher un mot, un couple de mot, un triplet de mots, etc.
  - Sur n'importe quelle ligne d'un tier CLAN (lignes principale et dépendantes)
  - Permet d'afficher les lignes avant ou après

Le trameur

- Logiciel de textométrie (Serge Fleury)
  - U. Paris 3 Sorbonne Nouvelle – CLA2T
  - http://tal.univ-paris3.fr/trameur/
- voir aussi Lexico3 (Lamalle, Martinez, Fleury, Salem)
  - http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/

Type de données utilisables

- Texte en vrac (issu de CLAN/ELAN ou nettoyé)

  *CHI: non !
  *CHI: allô !
  *CHI: là !
  *CHI: ça va ?
  *CHI:

- Texte avec des sections

  ww@fs porte . §
  les poubelles . §
  ww@fs voitures ! §
  ww@fs yy +... §
  a@fs voiture ! §

- Texte avec une structure (cadre)

  <Child="antoine-18-1_05_21">
  . §
  <Child="antoine-25-1_11_04">
  xx ? §
  coincé ? §

Générer des fichiers pour les outils de textométrie

- Utiliser les corpus extraits à l'aide de CLAN ou ELAN
  - Par exemple, extrait à l'aide de **kwal** et **flo**
    - **kwal +d +f +tCHI antoine-34-2_03_05.cha**
    - **flo +d antoine-34-2_03_05.kwal.cex**

➔ **Texte en vrac !**

*CHI: non !
*CHI: .
*CHI: oui !
*CHI: oui e garde le manteau !
*CHI: garde le manteau !

Texte avec des sections

- Solution (1)
  - **kwal +d +f +tCHI antoine-34-2_03_05.cha**
  - **flo +d antoine-34-2_03_05.kwal.cex**
  - Modifier avec un véritable « éditeur » de fichiers texte
    - Editer avec Notepad++
- Solution (2)
  - Utiliser un outil de traitement de corpus: R

Texte avec une structure

- Solution (1)
  - **kwal +d +f +tCHI +o@G antoine-34-2_03_05.cha**
  - **flo +d +t@ antoine-34-2_03_05.kwal0.cex**
  - Modifier avec un véritable « éditeur » de fichiers texte
    - Editer avec Notepad++
- Solution (2)
  - Utiliser un outil de traitement de corpus: R

Le Trameur - Le Métier Lexicometrique @CLA2T-P3 V. 9.0b063

Cadre | Ventilation | Section | Forme-Lemme | Catégorie-Tag | Segment | Cooc | Spécif | Concordance | Patron | Graphe | Sélection | Rapport | Param

Ventilation d'item :

[sélection po] RegExp

(Ecrire motif supra puis <Entrée>)

Sélection Annotation :

Forme Lemme Catégorie

Annotation sélectionnée :

1 Forme

Type de graphe :

Bar Ligne

Parties

G

(Ctrl-Clic : désélection partie)

Paramètres :

Taille de l'image

800x600 1000x800

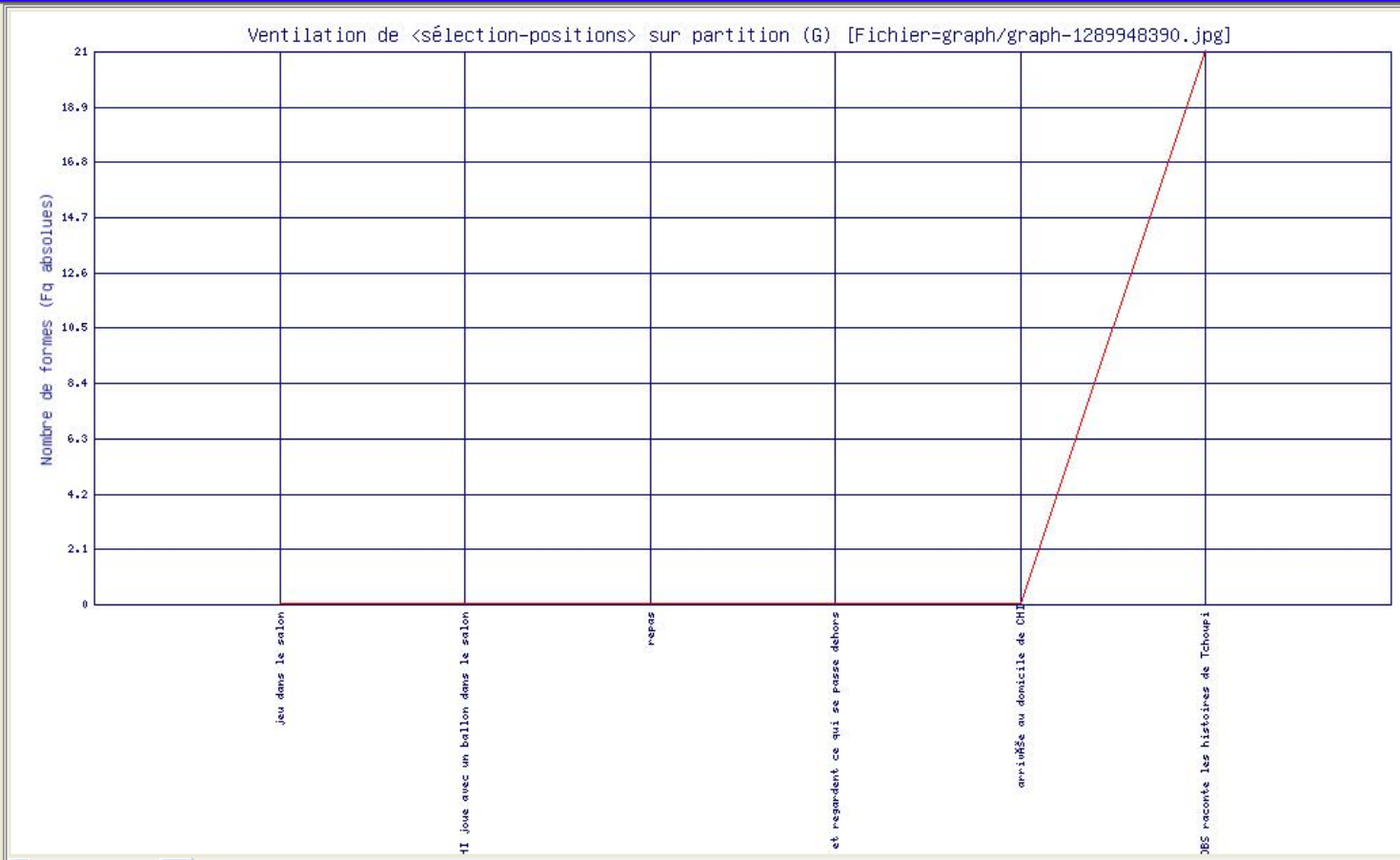
Fréquence

Absolues Relatives Spécifs

Accroissement de vocabulaire :

Trace Trame

Ajouter au Rapport :



Chargement de la partie : G

<Jeu de positions : {positions(Tchoupi)} >
<arrivée au domicile de CHI>
Fréquence de {sélection-positions(Tchoupi)} : 0
<CHI et OBS sont sur le balcon et regardent ce qui se passe dehors>
Fréquence de {sélection-positions(Tchoupi)} : 0
<CHI joue avec un ballon dans le salon>
Fréquence de {sélection-positions(Tchoupi)} : 0
<jeu dans le salon>
Fréquence de {sélection-positions(Tchoupi)} : 0
<OBS raconte les histoires de Tchoupi>
Fréquence de {sélection-positions(Tchoupi)} : 21
</corpus>

Texte avec une structure : fichiers multiples

- Solution (1)
  - **kwal +u +d +f +tCHI +o@G +o@Media antoine*.cha**
  - **flo +d +t@ antoine-tous.kwal.cex**
  - Modifier avec un véritable « éditeur » de fichiers texte
  - Editer avec Notepad++
- Solution (2)
  - Utiliser un outil de traitement de corpus: R

Le Trameur - Le Métier Lexicometrique @CLA2T-P3 V. 9.0b063

Cadre | Ventilation | Section | Forme-Lemme | Catégorie-Tag | Segment | Cooc | Spécif | Concordance | Patron | Graphe | Sélection | Rapport | Param

Ventilation d'item :

(sélection po | RegExp

(Ecrire motif supra puis <Entrée>)

Sélection Annotation :

Forme | Lemme | Catégorie

Annotation sélectionnée :

1 Forme

Type de graphe :

Bar | Ligne

Parties

Enfant

G

(Ctrl-Clic : désélection partie)

Paramètres :

Taille de l'image

800x600 | 1000x800

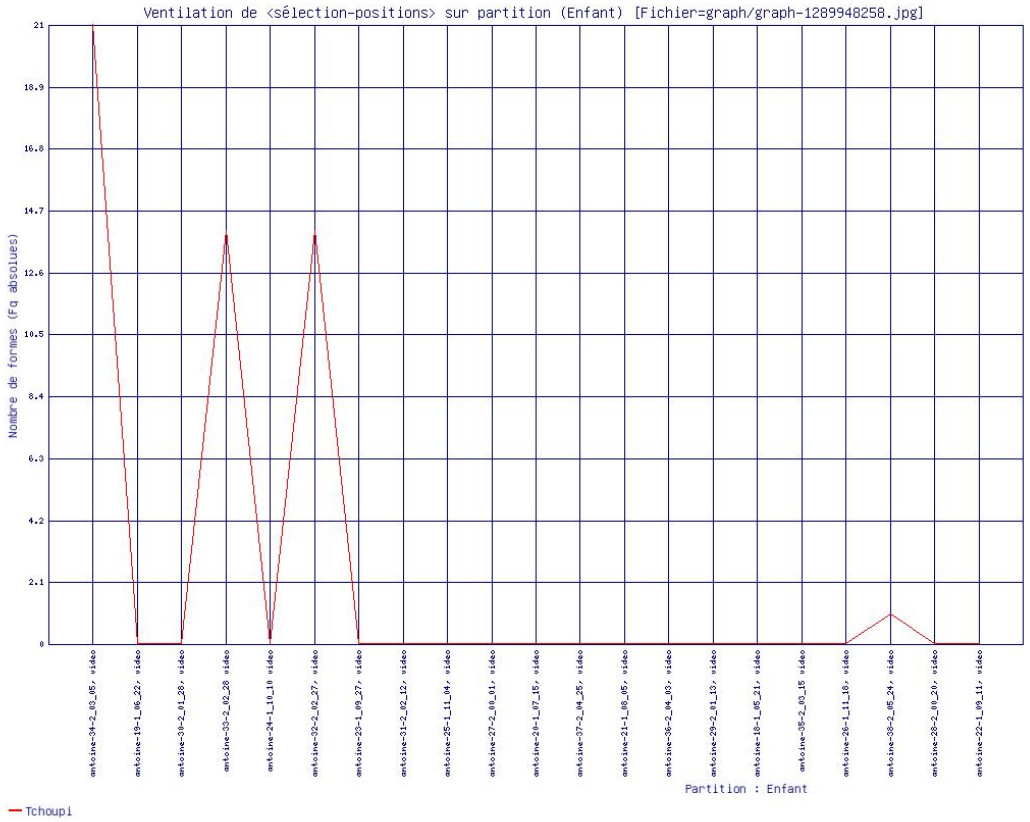
Fréquence

Absolues | Relatives | Spécifs

Accroissement de vocabulaire :

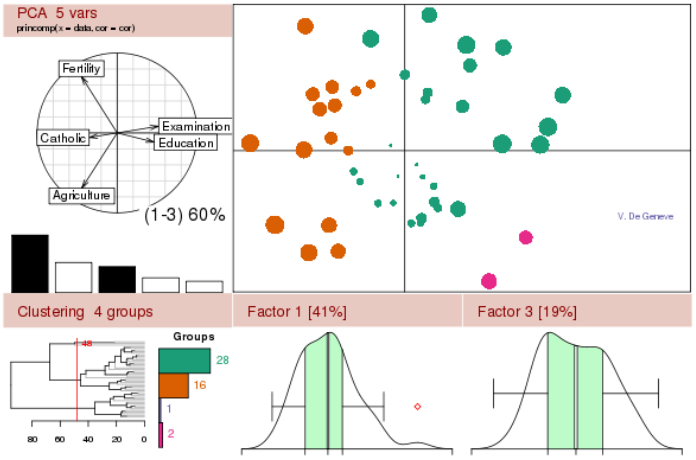
Trace Trame

Ajouter au Rapport :



Chargement de la partie : Enfant

<Jeu de positions : {positions(Tchoupi)} >
<antoine-18-1_05_21, video>
Fréquence de {sélection-positions(Tchoupi)} : 0
<antoine-19-1_06_22, video>
Fréquence de {sélection-positions(Tchoupi)} : 0
<antoine-20-1_07_15, video>



R

- Logiciel de statistique et de manipulation de fichiers en format texte
  - Manipuler, mettre en forme les données statistiques
    - Créer, modifier, éditer, des tables
      - des lignes, des cases
        - » en format texte, en format numérique
  - Réaliser des calculs statistiques
- Logiciel gratuit open source multiplateforme

- Préparation minimale des fichiers
  - flo +d antoine*.cha
  - longtier antoine*.flo.cex
    - Sur une seule ligne sans signes spéciaux
- Fichier exemple.r
  - rm(list=ls(all=T))
  - setwd("c:/abc/workshop")
- Chargement d'un fichier en mémoire
  - textbrut <- scan("antoine-38-2_05_24.flo.longtr.cex", what="char", sep="\n", quote="", encoding="UTF-8")
- Traitement du fichier
  - chi <- textbrut[grep("*CHI:\t", textbrut)]
  - lignep <- gsub(".*\t", "", chi)
  - enfant <- paste(lignep, "§")
  - enfant <- c("<enfant=\"2_05_24\">", enfant)
- Écriture du résultat
  - cat(enfant, file="chi.txt", sep="\n")

Répéter les traitements

- Transformer exemple en fonction

```
fichier <- function ( nomfic, resultat ) {
        textbrut <- scan( nomfic, what="char", sep="\n", quote="", encoding="UTF-8")
        chi <- textbrut[grep("*CHI:\t", textbrut)]
        lignep <- gsub( ".*\t", "", chi)
        enfant <- paste( lignep, "§" )
        cat( enfant, file="tmp", sep="\n" )
        f <- scan( "tmp", what="char", sep="\n", quote="")
        f <- gsub( "<.*>", "ww", f, perl=T)
        f <- c( paste( "<enfant=\"", nomfic, "\">", sep=""), f)
        cat( f, file=resultat, sep="\n", append=T )
}
```

- Utiliser sur un ensemble de fichiers

```
fa <- list.files(pattern="antoine.*flo.longtr.cex")

unlink("res.txt")
for (i in 1:length(fa)) {
    fichier( fa[i], "res.txt" )
}
```

Complexifier pour obtenir des formats plus complexes : MkAlign

- MkAlign permet de reproduire le travail des outils textométriques sur deux fichiers parallèles
  - Usage intéressant en traduction
- Mettre en parallèle les productions de l'enfant et des adultes
  - Transformer l'enfant et l'adulte sur la base de leurs tours de parole

*MOT: tu proposes à Martine un café ?
---
*CHI: tu veux un café ?
---
*OBS: oh oui volontiers !
*OBS: merci .
*MOT: xx café .
---
*CHI: on va faire tous les deux .
*CHI: tu veux un café ?
---
*OBS: oh oui s'il+te+plaît .
*OBS: qu' est+ce+que tu portes à la main ?
---
*CHI: hm xx hm des fleurs .
---
*OBS: des fleurs ?
---
*CHI: de fleurs ça !
---
*OBS: oui .
*OBS: elles tournent avec le vent .
---
*CHI: .

- <Child="madeleine-pol-14-2_02_06">
- . §
- tu veux un café ? §
- on va faire tous les deux . tu veux un café ? §
- hm xx hm des fleurs . §
- de fleurs ça ! §
- . §

- <Adults="madeleine-pol-14-2_02_06">
- tu proposes à Martine un café ? §
- oh oui volontiers ! merci . xx café . §
- oh oui s'il+te+plaît . qu' est+ce+que tu portes à la main ? §
- des fleurs ? §
- oui . elles tournent avec le vent . §