Le projet Les Vocaux : bilan d'étape

Julie Glikman*,**, Nicolas Mazziotta***,
Camille Fauth*, Christophe Benzitoun**

*Université de Strasbourg, LiLPa (France)
**Université de Lorraine, ATILF (France)
***Université de Liège, Traverses (Belgique)

Sciences participatives et nouvelles données
Nancy, 30 sept. 2022

Les VOCAUX

Introduction

Qu’est-ce qu’un « vocal » ?

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :
  - Large variété de plateformes de communication les prenant en charge (Snapchat, Messenger, Whatsapp, etc., mais également messagerie standard)

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :
  - Large variété de plateformes de communication les prenant en charge (Snapchat, Messenger, Whatsapp, etc., mais également messagerie standard)
  - **Populaires** et relayés dans la presse

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :
  - Large variété de plateformes de communication les prenant en charge (Snapchat, Messenger, Whatsapp, etc., mais également messagerie standard)
  - **Populaires** et relayés dans la presse
  - Se substituent à l'envoi de SMS et aux appels

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :
  - Large variété de plateformes de communication les prenant en charge (Snapchat, Messenger, Whatsapp, etc., mais également messagerie standard)
  - **Populaires** et relayés dans la presse
  - Se substituent à l'envoi de SMS et aux appels
- Exemple : LesVocaux, 02-01

Introduction

Qu'est-ce qu'un « vocal » ?

- Concept de « SMS vocal », « message vocal », « note vocale » ou « vocal » :
  - Large variété de plateformes de communication les prenant en charge (Snapchat, Messenger, Whatsapp, etc., mais également messagerie standard)
  - **Populaires** et relayés dans la presse
  - Se substituent à l'envoi de SMS et aux appels
- Exemple : LesVocaux, 02-01
  - euh la vidéo . . . où v- euh . . . on est en accéléré elle est hyper drôle . . . (LesVocaux, 02-01)

Introduction

Caractéristiques communicatives

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,
- En exploitant Koch et Oesterreicher (2001) :

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,
- En exploitant Koch et Oesterreicher (2001) :
  - Code oral

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,
- En exploitant Koch et Oesterreicher (2001) :
  - Code oral
  - **Proximité** communicative ($\neq$ publications Tiktok/instagram) : communication privée, interlocuteur connu (pas nécessairement intime), globalement spontanée, émotionnalité potentiellement élevée

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,
- En exploitant Koch et Oesterreicher (2001) :
  - Code oral
  - **Proximité** communicative ($\neq$ publications Tiktok/instagram) : communication privée, interlocuteur connu (pas nécessairement intime), globalement spontanée, émotionnalité potentiellement élevée
  - **Distance** communicative ($\neq$ communication téléphonique) : monologues, coopération communicative faible (pas de chevauchement, ni de coproduction), généralement pas de coprésence spacio-temporelle

Introduction

Caractéristiques communicatives

- langue **spontanée** dans un contexte de production non dirigée, aux échanges de type privés, non surveillés,
- En exploitant Koch et Oesterreicher (2001) :
  - Code oral
  - **Proximité** communicative ($\neq$ publications Tiktok/instagram) : communication privée, interlocuteur connu (pas nécessairement intime), globalement spontanée, émotionnalité potentiellement élevée
  - **Distance** communicative ($\neq$ communication téléphonique) : monologues, coopération communicative faible (pas de chevauchement, ni de coproduction), généralement pas de coprésence spacio-temporelle

$\rightarrow$ Catégorie **spécifique**

Introduction

Type de communication peu étudié + besoin

- Multiplication des corpus oraux (`https://www.ortolang.fr/`)

Introduction

Type de communication peu étudié + besoin

- Multiplication des corpus oraux (`https://www.ortolang.fr/`)
- SMS bien étudiés (*sms4science*, *What's up Switzerland*)

Introduction

Type de communication peu étudié + besoin

- Multiplication des corpus oraux (`https://www.ortolang.fr/`)
- SMS bien étudiés (*sms4science*, *What's up Switzerland*)
- Peu d'études sur les vocaux (exceptions rares, comme Estade et le Maire 2018)

Introduction

Type de communication peu étudié + besoin

- Multiplication des corpus oraux (`https://www.ortolang.fr/`)
- SMS bien étudiés (*sms4science*, *What's up Switzerland*)
- Peu d'études sur les vocaux (exceptions rares, comme Estade et le Maire 2018)
- Parole spontanée naturelle toujours **difficile d'accès** (récolte sous forme d'entretiens)

Introduction

Objectif de l'exposé

- Présentation des méthodes de recueil, et de construction du corpus
- Bilan sur le recueil
- Premières observations (tests sur un corpus « martyr »)

Plan

Recueil des données

Recueil des données

Nature technique des énonciations

- En tant qu'actes d'énonciation, ce sont des messages **enregistrés**
- La récolte peut commencer par un simple transfert

Recueil des données

Nature technique des énonciations

- En tant qu'actes d'énonciation, ce sont des messages **enregistrés**
- La récolte peut commencer par un simple transfert

Deux étapes de recueil des données

1. recueil expérimental préalable : « Martyr »
2. recueil à grande échelle : « Campagne 2022 »

Recueil des données

Martyr

- 7 participantes volontaires
- Entretiens individuels sur les pratiques
- Partage des vocaux (sept.-nov. 2021)
- Recueil échantillon d'entretiens (matériaux contrastifs)

Recueil des données

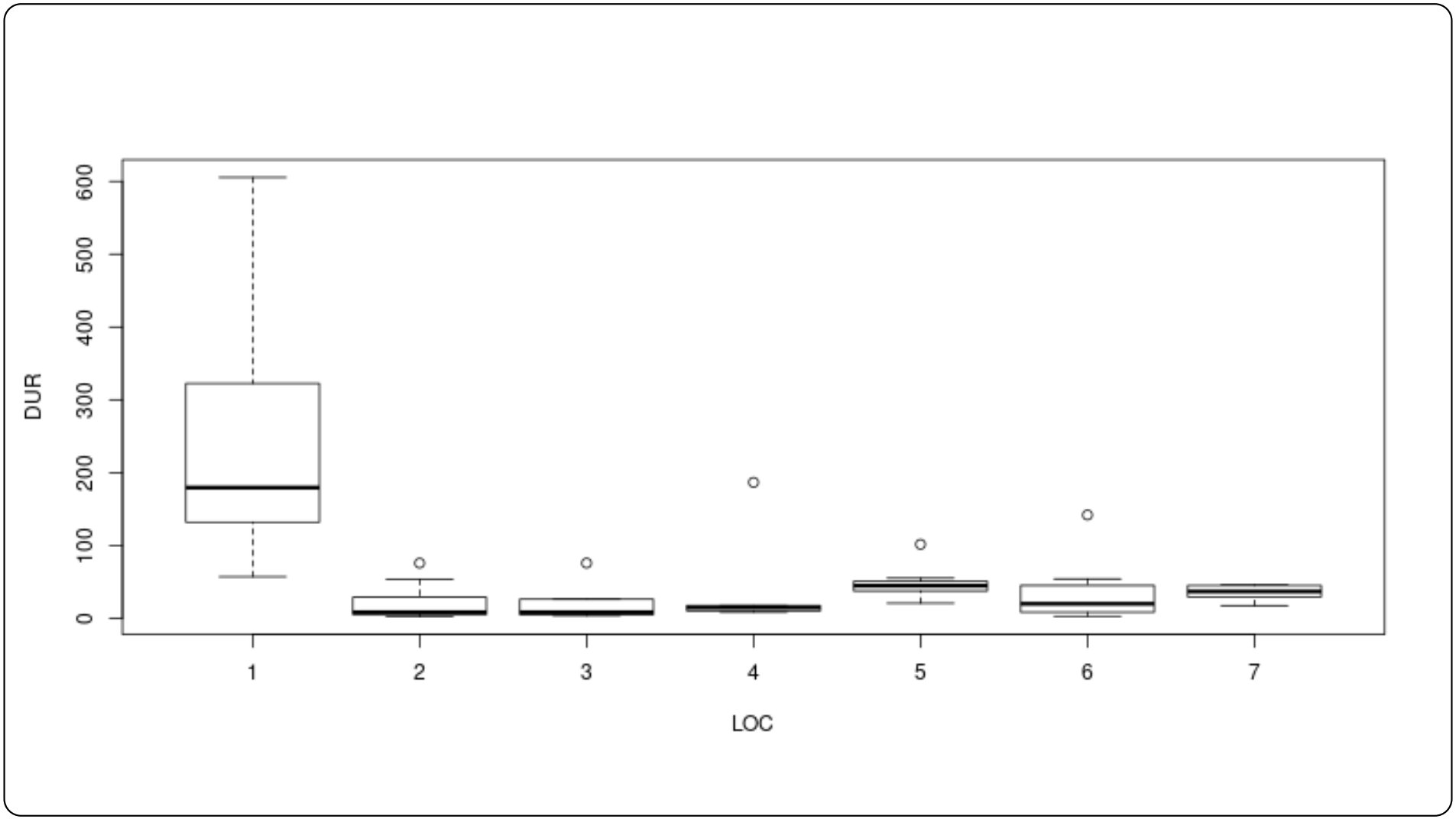
Martyr

- 7 participantes volontaires
- Entretiens individuels sur les pratiques
- Partage des vocaux (sept.-nov. 2021)
- Recueil échantillon d'entretiens (matériaux contrastifs)

Données recueillies

- Nombre de locutrices : 7
- Origine : Strasbourg et Paris
- Nombre de vocaux : 66
- Durée totale : 1h00m44.73s
- Une distribution variable des durées (cas atypique)

Recueil des données



Distribution des durées par locuteur

Recueil des données

Campagne 2022

- **Crowdsourcing** (cp. *Français de nos régions* et *Donnez votre français à la science* (Glikman et al.)

Recueil des données

Campagne 2022

- **Crowdsourcing** (cp. *Français de nos régions* et *Donnez votre français à la science* (Glikman et al.)
- Nécessité de respecter la RGPD (voix = donnée sensible)

Recueil des données

Campagne 2022

- **Crowdsourcing** (cp. *Français de nos régions* et *Donnez votre français à la science* (Glikman et al.)
- Nécessité de respecter la RGPD (voix = donnée sensible)
- Questionnaire en ligne (métadonnées + habitudes d'utilisation)

Recueil des données

Campagne 2022

- **Crowdsourcing** (cp. *Français de nos régions* et *Donnez votre français à la science* (Glikman et al.)
- Nécessité de respecter la RGPD (voix = donnée sensible)
- Questionnaire en ligne (métadonnées + habitudes d'utilisation)
- Campagne de recueil limitée dans le temps (avril – septembre 2022)

Recueil des données

Campagne 2022

- **Crowdsourcing** (cp. *Français de nos régions* et *Donnez votre français à la science* (Glikman et al.)
- Nécessité de respecter la RGPD (voix = donnée sensible)
- Questionnaire en ligne (métadonnées + habitudes d'utilisation)
- Campagne de recueil limitée dans le temps (avril – septembre 2022)
- Campagne de publicité ayant surtout attiré des **étudiants et collègues** (Facebook, Instagram, annonces aux cours)

Recueil des données

Métadonnées

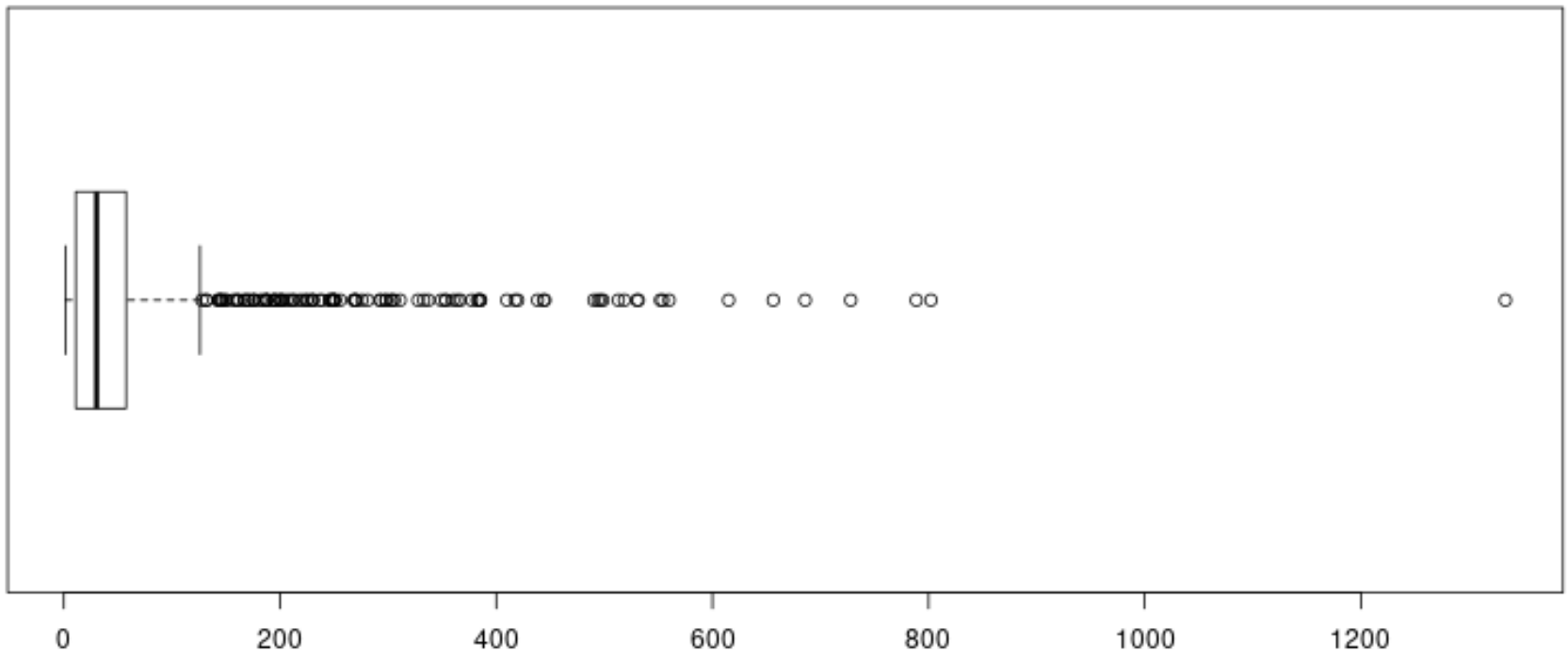
- speaker : numéro d'identification du locuteur
- gender : genre
- age : âge
- current location : lieu de vie actuel
- current country : pays où se trouve le lieu de vie
- home location : ville natale
- home country : pays natal
- message usage : fréquence d'utilisation des vocaux
- filename : nom du fichier (référence du texte)

Recueil des données

Données recueillies

- Nombre de locuteurs/ices : 45
- Origine : France, Belgique, Suisse, Canada (1)
- Nombre de vocaux : 1160
- Durée totale : 19h06m49.2s
- Une distribution variable des durées (non évaluées par locuteur)

Recueil des données



Distribution des durées

Recueil des données

Données recueillies

- Nombre de locuteurs/ices : 45
- Origine : France, Belgique, Suisse, Canada (1)
- Nombre de vocaux : 1160
- Durée totale : 19h06m49.2s
- Une distribution variable des durées (non évaluées par locuteur)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.313	10.992	30.241	59.318	57.561	1333.678

Recueil des données

Données recueillies

- Nombre de locuteurs/ices : 45
- Origine : France, Belgique, Suisse, Canada (1)
- Nombre de vocaux : 1160
- Durée totale : 19h06m49.2s
- Une distribution variable des durées (non évaluées par locuteur)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.313	10.992	**30.241**	59.318	57.561	**1333.678**

Premières observations

Premières observations

1re approche « intuitive » : types discursifs

Des types discursifs variés :

Premières observations

1re approche « intuitive » : types discursifs

Des types discursifs variés :

- **Discours/Narration** : vingt heures cinquante-sept je sors à peine du . . . du magasin . . . parce que du coup euh la caisse ne marchait pas donc j'ai dû tout retaper à la main . . . et en fait euh il y avait un souci parce que [. . .] (LesVocaux, 05-06)

Premières observations

1re approche « intuitive » : types discursifs

Des types discursifs variés :

- **Discours/Narration** : vingt heures cinquante-sept je sors à peine du . . . du magasin . . . parce que du coup euh la caisse ne marchait pas donc j'ai dû tout retaper à la main . . . et en fait euh il y avait un souci parce que [. . .] (LesVocaux, 05-06)
- **Interactions** : bon ben tant pis ça y est tu fais la gueule ou quoi (LesVocaux, 08-02)

Premières observations

1re approche « intuitive » : types discursifs

Des types discursifs variés :

- **Discours/Narration** : vingt heures cinquante-sept je sors à peine du . . . du magasin . . . parce que du coup euh la caisse ne marchait pas donc j'ai dû tout retaper à la main . . . et en fait euh il y avait un souci parce que [. . .] (LesVocaux, 05-06)
- **Interactions** : bon ben tant pis ça y est tu fais la gueule ou quoi (LesVocaux, 08-02)
- **Discours cité** : il sort à #name euh . . . ouais euh vous pensez à bien garder les distances de sécurité et euh alors elle dit bah oui vous voyez bien qu'ils sont espacés . . . oui oui non mais je vois je vois hein (LesVocaux, 05-08)

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

- Particulièrement fréquents dans le corpus

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

- Particulièrement fréquents dans le corpus
- Typiques de l'oral

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

- Particulièrement fréquents dans le corpus
- Typiques de l'oral
- Peu normés (non conformes aux descriptions normées)

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

- Particulièrement fréquents dans le corpus
- Typiques de l'oral
- Peu normés (non conformes aux descriptions normées)
- Peu étudiés (ou à explorer davantage)

Premières observations

1re approche « intuitive » : phénomènes saillants

Nous pouvons observer des phénomènes :

- Particulièrement fréquents dans le corpus
- Typiques de l'oral
- Peu normés (non conformes aux descriptions normées)
- Peu étudiés (ou à explorer davantage)

$\Rightarrow$ Autant de **constructions**, **emplois de mots** et de **phénomènes prosodiques** que nous allons évoquer

Premières observations

Structures interrogatives

Structures **fréquentes** dans le corpus (dont interrogatives *in situ*, en *est-ce que*,. . .) – Coveney 2020

Premières observations

Structures interrogatives

Structures **fréquentes** dans le corpus (dont interrogatives *in situ*, en *est-ce que*,. . .) – Coveney 2020

- alors je dois faire comment ? (LesVocaux, 02-07)
- ouais comment on peut s'organiser du coup ? (LesVocaux, 07-02)
- tu fais la gueule ou quoi ? (LesVocaux, 06-08)
- est-ce que nous on peut prendre un train jusqu'à euh ... jusqu'à #city jusqu'à #city ? (LesVocaux, 02-05)
- qu'est-ce qui t'a déclenché cette pensée ? (LesVocaux, 06-03)

Premières observations

Pseudo-clivées
Également **typiques** de l'oral – Roubaud 2000

Premières observations

Pseudo-clivées

Également **typiques** de l'oral – Roubaud 2000

- euh ce qu'on pourrait faire mardi pro du coup ... c'est que ... euh ... on pourrait euh aller manger une glace en ville (LesVocaux, 07-06)
- comme si euh comme si le seul ... le seul truc ... qui était ennuyeux dans cette affaire ... euh c'est qu'il ait énervé sa mère (LesVocaux, 01-03)

Premières observations

meuf en apostrophe

À intégrer à la problématique des **termes d'adresse** (y compris question de la diachronie courte) – Bruno 2013, Lagorgette 2003, Lagorgette 2006

Premières observations

meuf en apostrophe

À intégrer à la problématique des **termes d'adresse** (y compris question de la diachronie courte) – Bruno 2013, Lagorgette 2003, Lagorgette 2006

- ... meuf ... #rire ... je viens de relire ... que vendredi soir j'avais euh j'avais appelé plein de fois (LesVocaux 06-04)
- je suis trop dég qu'ils aient enlevé How I meet your mother de ... de Netflix meuf (LesVocaux, 02-06)
- juste ... meuf ... histoire hyper drôle genre j'étais soule au lieu de faire un virement de cinq euro j'ai fait un virement de cinq cent euro (LesVocaux, 02-12)

Premières observations

meuf en apostrophe

À intégrer à la problématique des **termes d'adresse** (y compris question de la diachronie courte) – Bruno 2013, Lagorgette 2003, Lagorgette 2006

- ... meuf ... #rire ... je viens de relire ... que vendredi soir j'avais euh j'avais appelé plein de fois (LesVocaux 06-04)
- je suis trop dég qu'ils aient enlevé How I meet your mother de ... de Netflix meuf (LesVocaux, 02-06)
- juste ... meuf ... histoire hyper drôle genre j'étais soule au lieu de faire un virement de cinq euro j'ai fait un virement de cinq cent euro (LesVocaux, 02-12)

meuf pas évalué dans la thèse de Wang (2016) sur le paradigme des appellatifs comme *Mademoiselle*

Premières observations

(In)subordination, (in)dépendance

parce que P non causales microsyntaxiquement indépendantes – Debaisieux 1994

- mais bref je voulais savoir si ... hum ... tu trouvais ça ... chouette parce qu'en fait j'ai de la peinture à doigts ... qu'on devait utiliser pour l'anniversaire à #name (LesVocaux, 05-04)
- mais en tout cas mais vraiment montez parce que c'est vraiment trop beau (LesVocaux, 02-07)

Premières observations

(In)subordination, (in)dépendance

parce que P non causales microsyntaxiquement indépendantes – Debaisieux 1994

- mais bref je voulais savoir si ... hum ... tu trouvais ça ... chouette parce qu'en fait j'ai de la peinture à doigts ... qu'on devait utiliser pour l'anniversaire à #name (LesVocaux, 05-04)
- mais en tout cas mais vraiment montez parce que c'est vraiment trop beau (LesVocaux, 02-07)

Autres formes dont il faut **évaluer le paradigme** : *même que*, *sauf que*, *alors que P* non temporelle, etc.

Premières observations

Marqueurs discursifs propositionnels

Question des « recteurs faibles » et de la parataxe – Andersen 2007, Bolly 2010, Galatanu et al. 2014

Premières observations

Marqueurs discursifs propositionnels

Question des « recteurs faibles » et de la parataxe – Andersen 2007, Bolly 2010, Galatanu et al. 2014

- j'avais j'avais besoin de réconfort tu vois (LesVocaux, 03-03)
- c'était chelou tu vois (LesVocaux, 06-06)
- euh je t'explique ... euh ... pour ... donc quand on va à Lyon ... pour revenir genre euh ... les trains ils sont ... enfin ils sont chers genre ça fait quand même genre ... euh cinquante balles (LesVocaux, 02-05)
- en fait je t'explique nous on avait garé la caisse là ... on avait fait un tour ... du petit centre-ville du port (LesVocaux, 02-07)

Premières observations

Marqueurs discursifs propositionnels

Question des « recteurs faibles » et de la parataxe – Andersen 2007, Bolly 2010, Galatanu et al. 2014

- j'avais j'avais besoin de réconfort tu vois (LesVocaux, 03-03)
- c'était chelou tu vois (LesVocaux, 06-06)
- euh je t'explique ... euh ... pour ... donc quand on va à Lyon ... pour revenir genre euh ... les trains ils sont ... enfin ils sont chers genre ça fait quand même genre ... euh cinquante balles (LesVocaux, 02-05)
- en fait je t'explique nous on avait garé la caisse là ... on avait fait un tour ... du petit centre-ville du port (LesVocaux, 02-07)

Pas de travaux sur la forme *je t'explique* (?)

Premières observations

Syntaxe : reformulations en « piles »

Blanche-Benveniste C. et al. 1979, Kahane et Pietrandrea 2012

- et je trouve que c'est ... hyper bizarre en fait ... euh le mécanisme ... de la désobéissance ou de la ... ouais du ouais du de la désobéissance au sens large disons de pas faire ce qu'on te demande ... chez les enfants et surtout chez les enfants euh ... euh ... un peu tu vois qui sont plus pas des tous petits ... qui vraiment ont beaucoup de mal à résister à ... à ... à l'attrait d'une chose et qui comprennent pas forcément le sens des interdits ... ou ... ou des ordres en fait ... qu'on leur qu'on leur donne (LesVocaux, 01-03)

Premières observations

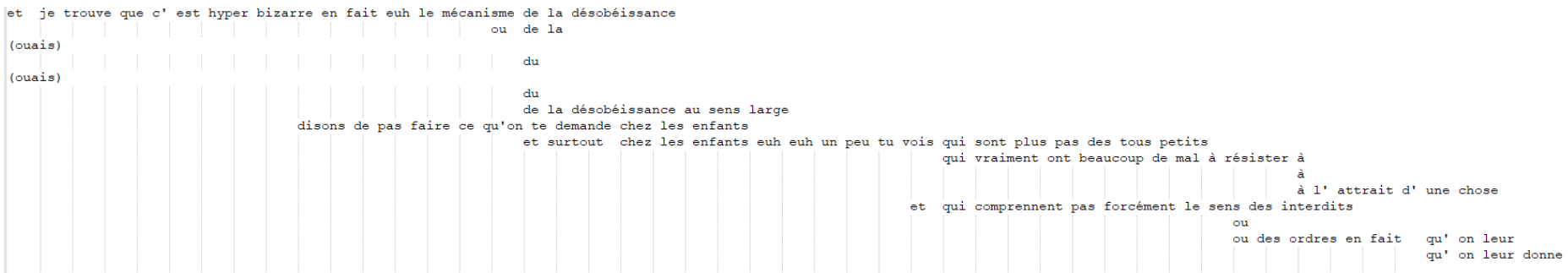
Syntaxe : reformulations en « piles »

Blanche-Benveniste C. et al. 1979, Kahane et Pietrandrea 2012

- et je trouve que c'est ... hyper bizarre en fait ... euh le mécanisme ... de la désobéissance ou de la ... ouais du ouais du de la désobéissance au sens large disons de pas faire ce qu'on te demande ... chez les enfants et surtout chez les enfants euh ... euh ... un peu tu vois qui sont plus pas des tous petits ... qui vraiment ont beaucoup de mal à résister à ... à ... à l'attrait d'une chose et qui comprennent pas forcément le sens des interdits ... ou ... ou des ordres en fait ... qu'on leur qu'on leur donne (LesVocaux, 01-03)

Nombreuses **piles imbriquées**, nombreux marqueurs de piles, nombreuses marques d'hésitation (pauses, *euh*)

Premières observations



Analyse en grille

Premières observations

Phonétique : question des pauses et des hésitations

- Nombreuses pauses (notées <. . . >) et hésitations

Premières observations

Phonétique : question des pauses et des hésitations

- Nombreuses pauses (notées <. . . >) et hésitations

$\Rightarrow$ Cela justifie qu'on s'intéresse :

- Aux **pauses** : interne/externe au syntagme
- Aux **allongements** : notamment leur positionnement par rapport aux piles notamment
- Aux ***fillers*** de type *euh*

Premières observations

Phonétique : question des pauses et des hésitations

- Nombreuses pauses (notées <. . . >) et hésitations

$\Rightarrow$ Cela justifie qu'on s'intéresse :

- Aux **pauses** : interne/externe au syntagme
- Aux **allongements** : notamment leur positionnement par rapport aux piles notamment
- Aux ***fillers*** de type *euh*

pour rendre compte des **stratégies de construction** de la chaîne phonique de l'énoncé

Premières observations

	Syl	Po	Pi	F	F+	F-	A	Dis/DT
Loc1	1296	82	84	8	66	6	6	24,60
Loc2	791	50	27	7	26	4	7	16,30
Loc3	230	8	7	0	10	0	1	14,01
Loc4	250	10	3	0	6	1	6	15,44
Loc5	1508	56	26	3	42	1	21	12,83
Loc6	1003	79	36	1	25	2	20	12,82
Loc7	1321	79	44	6	43	12	19	17,71
Totaux	6399	364	227	25	218	26	80	

Données phonétiques (Glikman et Fauth 2022)

Chaîne de traitement

Chaîne de traitement

Un corpus pour étudier ces phénomènes

- **Annotations** nécessaires :
  - Ph **phonétique** (phones, syllabation, pauses, alignement au signal)
  - Le **lemmatisation**
  - MS **morphosyntaxe** (PDD, morphologie, syntaxe dépendancielle)

Chaîne de traitement

Un corpus pour étudier ces phénomènes

- **Annotations** nécessaires :
  - Ph **phonétique** (phones, syllabation, pauses, alignement au signal)
  - Le **lemmatisation**
  - MS **morphosyntaxe** (PDD, morphologie, syntaxe dépendancielle)
- Permettront d'évaluer les **interférences** :
  - Distribution relative des pauses et des piles : Ph $\times$ MS
  - Typologie des allongements : Ph $\times$ Le et Ph $\times$ MS
  - Comportement syntaxique des apostrophes, marqueurs de discours, etc. : Le $\times$ MS

Chaîne de traitement

Un corpus pour étudier ces phénomènes

- **Annotations** nécessaires :
  - Ph **phonétique** (phones, syllabation, pauses, alignement au signal)
  - Le **lemmatisation**
  - MS **morphosyntaxe** (PDD, morphologie, syntaxe dépendancielle)
- Permettront d'évaluer les **interférences** :
  - Distribution relative des pauses et des piles : Ph $\times$ MS
  - Typologie des allongements : Ph $\times$ Le et Ph $\times$ MS
  - Comportement syntaxique des apostrophes, marqueurs de discours, etc. : Le $\times$ MS
- Y compris **interférences à « trois voies »**
  - Réalisation des marqueurs de discours : Le $\times$ Ph $\times$ MS
  - « Usure phonique » [SHi], [Sepa], liaisons etc. : Le $\times$ Ph $\times$ MS

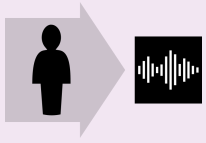
Chaîne de traitement

Un corpus pour étudier ces phénomènes

- **Annotations** nécessaires :
  - Ph **phonétique** (phones, syllabation, pauses, alignement au signal)
  - Le **lemmatisation**
  - MS **morphosyntaxe** (PDD, morphologie, syntaxe dépendancielle)
- Permettront d'évaluer les **interférences** :
  - Distribution relative des pauses et des piles : Ph $\times$ MS
  - Typologie des allongements : Ph $\times$ Le et Ph $\times$ MS
  - Comportement syntaxique des apostrophes, marqueurs de discours, etc. : Le $\times$ MS
- Y compris **interférences à « trois voies »**
  - Réalisation des marqueurs de discours : Le $\times$ Ph $\times$ MS
  - « Usure phonique » [SHi], [Sepa], liaisons etc. : Le $\times$ Ph $\times$ MS

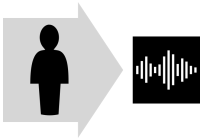
Distribuer le corpus dans des formats utilisables :
Praat, **CoNLL**, **Texte**, **XML pour TXM**,

Chaîne de traitement

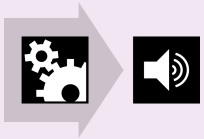


Vocal production (misc. platforms)

Chaîne de traitement

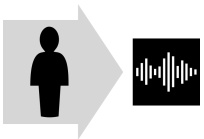


Vocal
production
(misc. platforms)

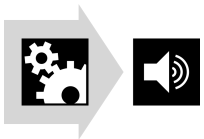


Conversion
to WAV
(audacity)

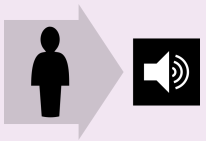
Chaîne de traitement



Vocal
production
(misc. platforms)

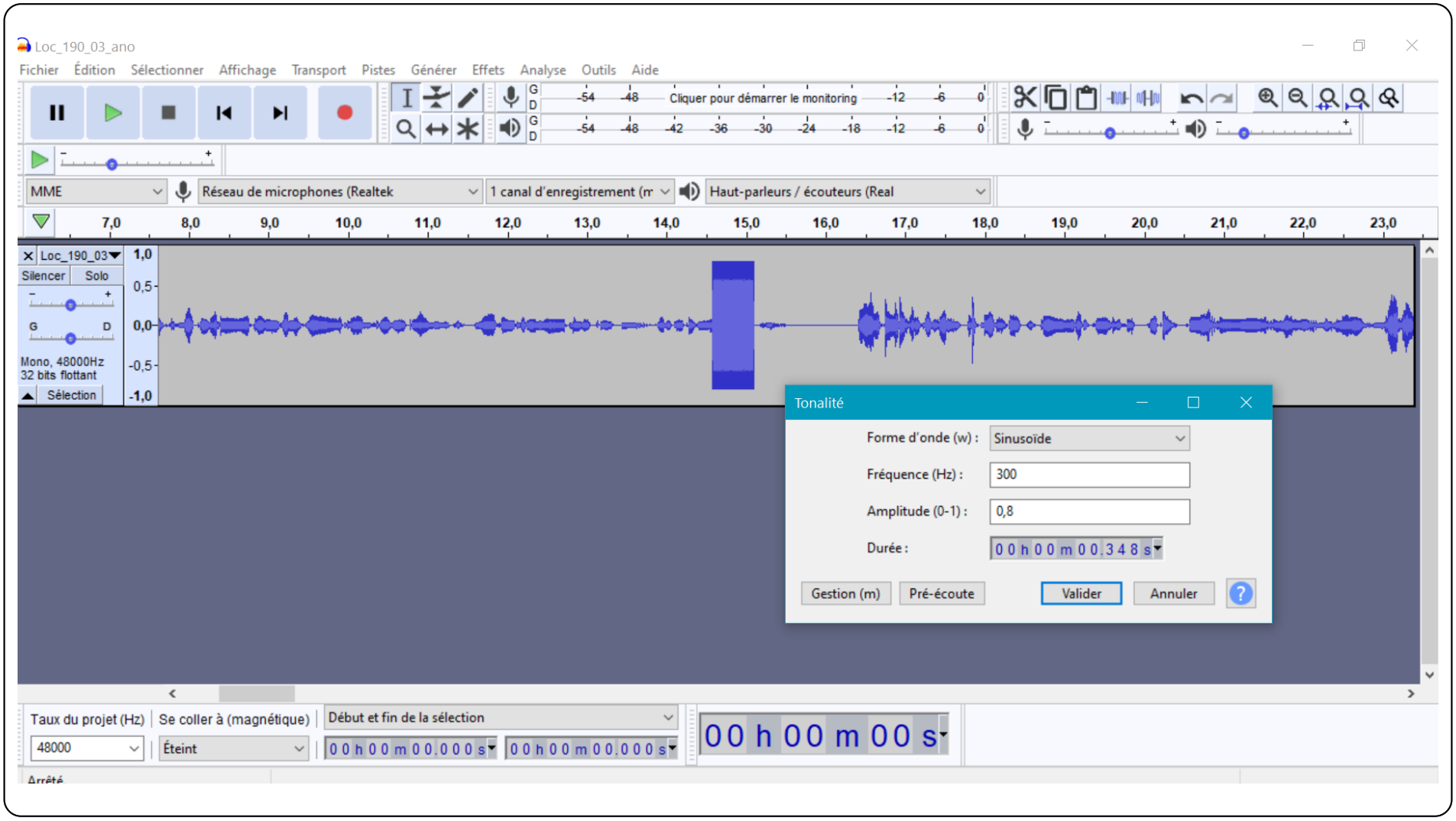


Conversion
to WAV
(audacity)



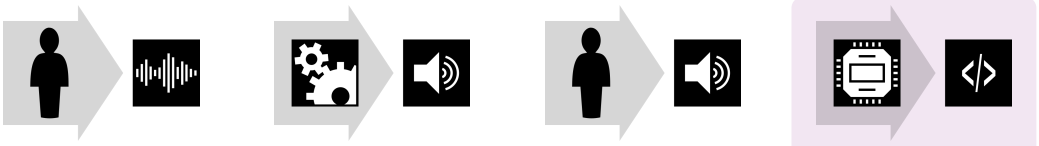
Signal
anonymization
(manual)

Chaîne de traitement



Anonymisation manuelle (Audacity)

Chaîne de traitement



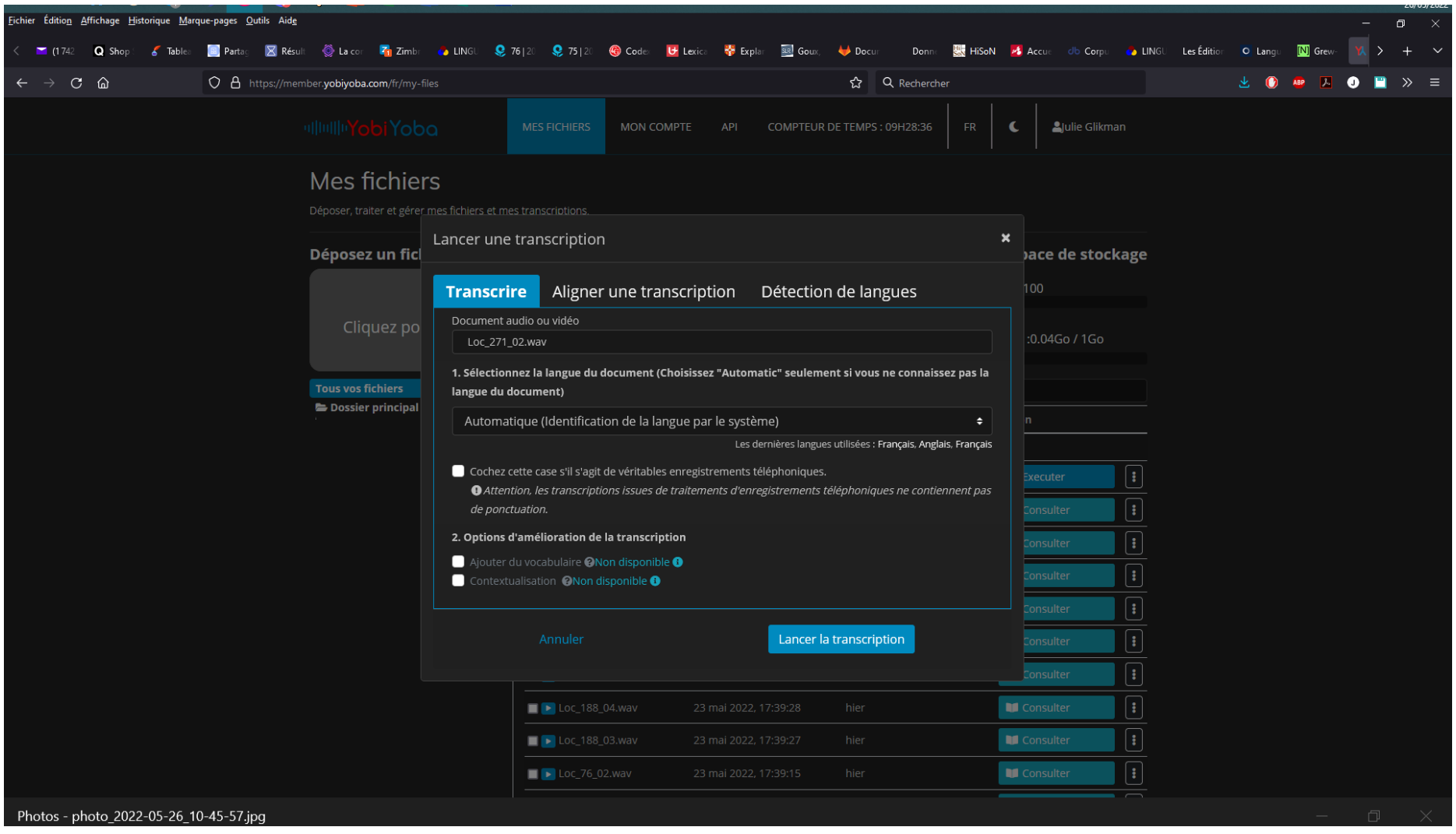
Vocal production (misc. platforms)

Conversion to WAV (audacity)

Signal anonymization (manual)

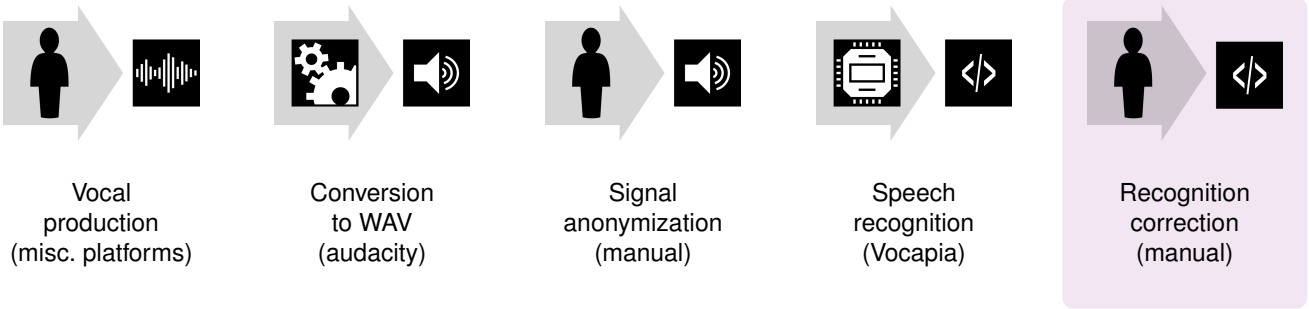
Speech recognition (Vocapia)

Chaîne de traitement

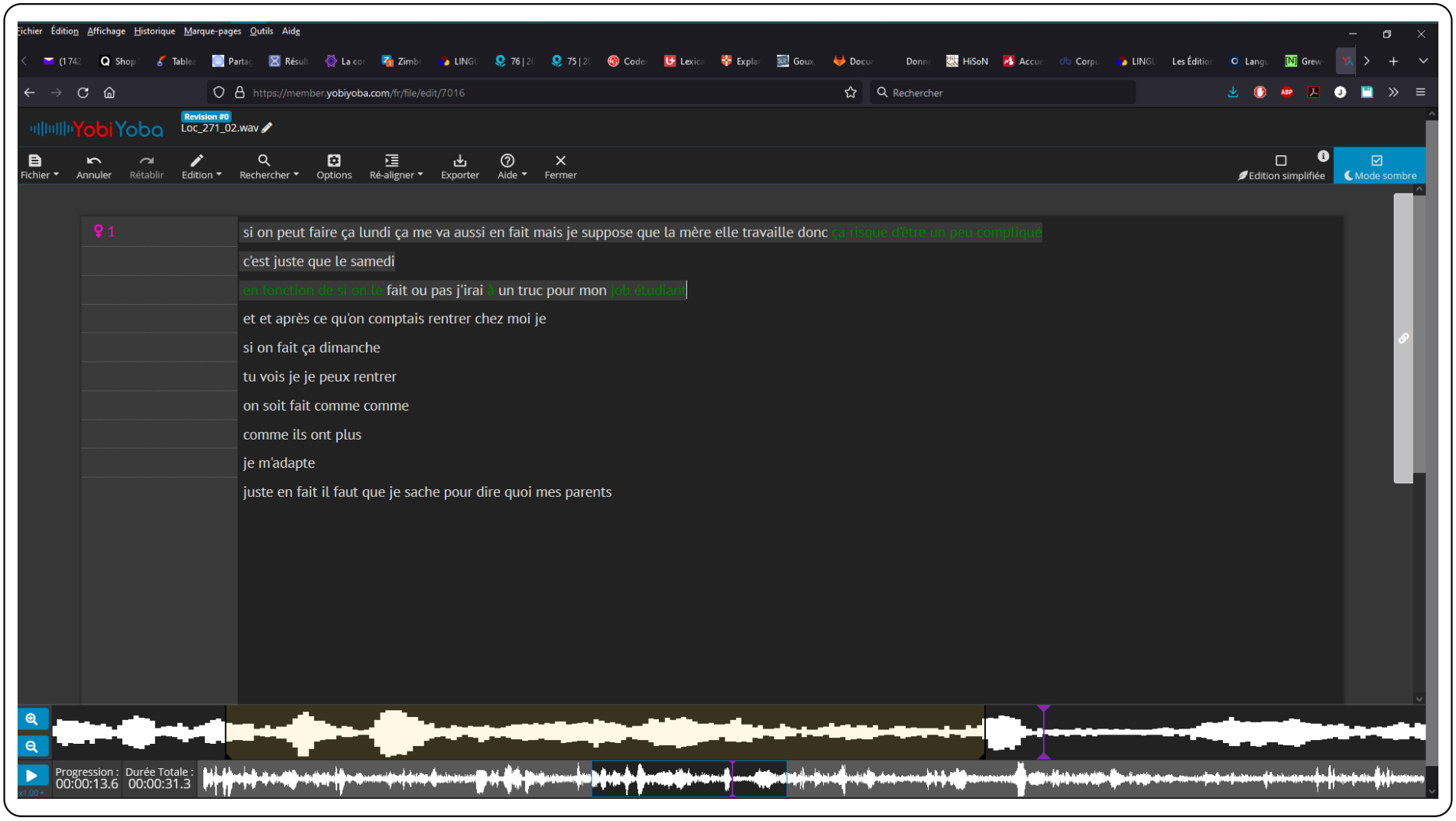


Transcription automatique (Vocapia sur Yobiyoba)

Chaîne de traitement

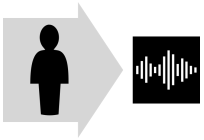


Chaîne de traitement

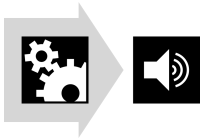


Correction manuelle (Vocapia)

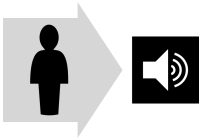
Chaîne de traitement



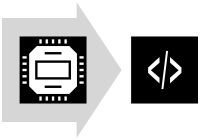
Vocal production (misc. platforms)



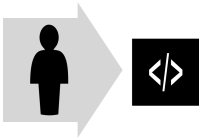
Conversion to WAV (audacity)



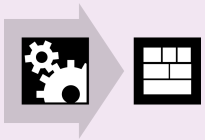
Signal anonymization (manual)



Speech recognition (Vocapia)

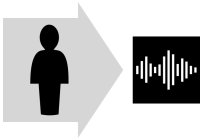


Recognition correction (manual)

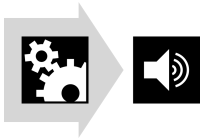


Conversion to Praat TextGrid (script)

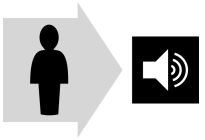
Chaîne de traitement



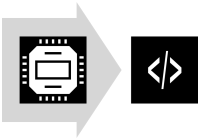
Vocal production (misc. platforms)



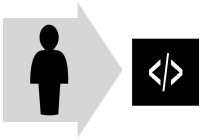
Conversion to WAV (audacity)



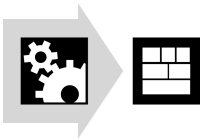
Signal anonymization (manual)



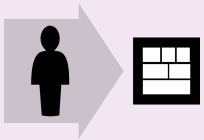
Speech recognition (Vocapia)



Recognition correction (manual)

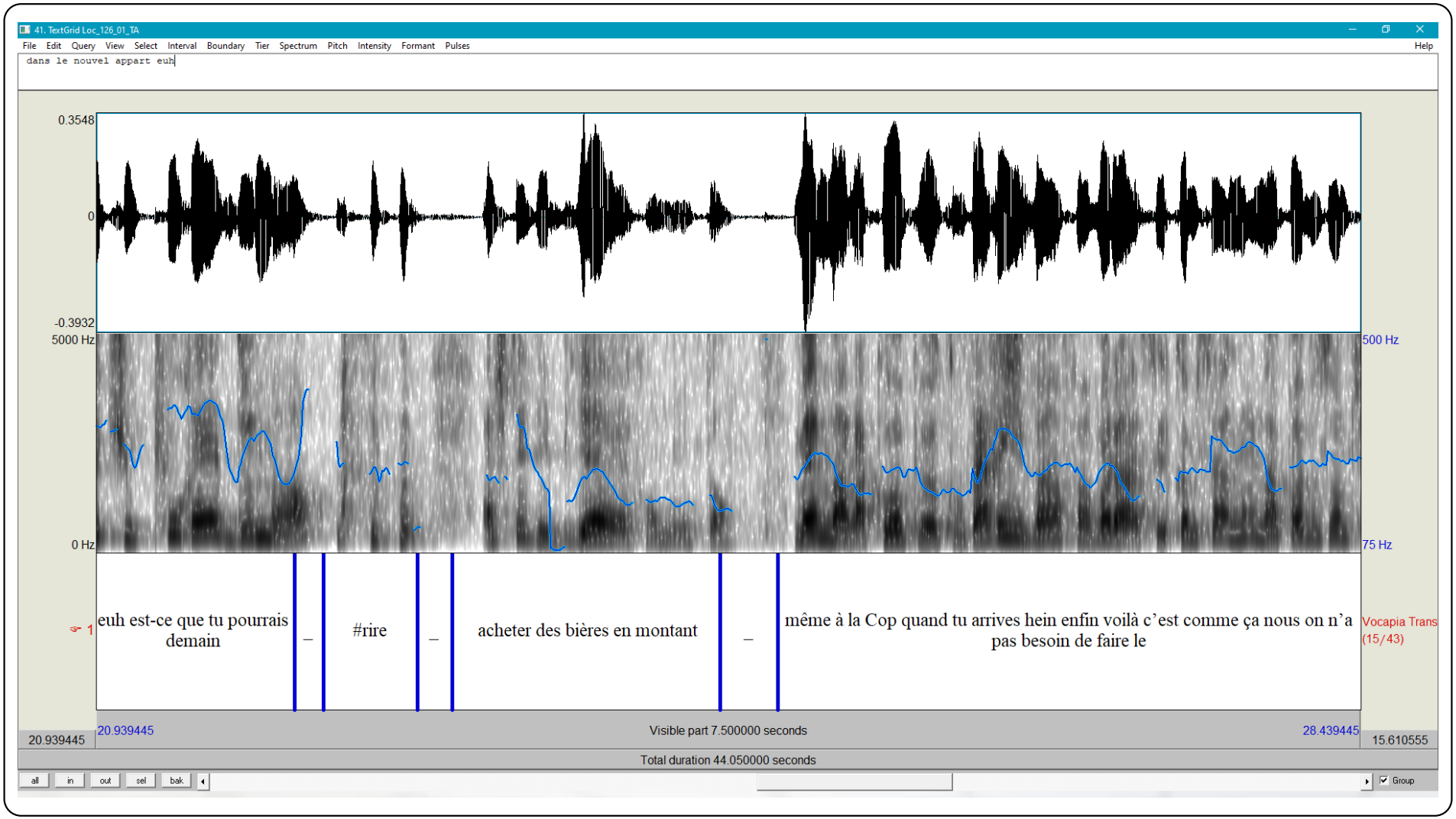


Conversion to Praat TextGrid (script)



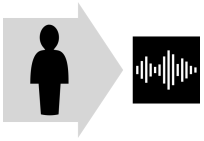
Phonetic analysis preparation (manual)

Chaîne de traitement

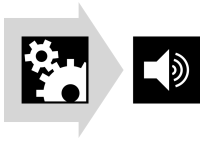


Préparation manuelle à l'analyse phonétique

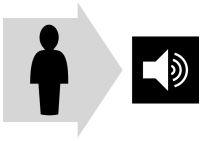
Chaîne de traitement



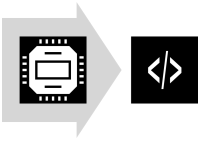
Vocal production (misc. platforms)



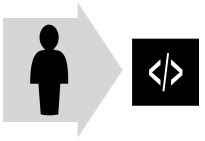
Conversion to WAV (audacity)



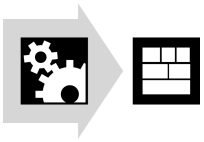
Signal anonymization (manual)



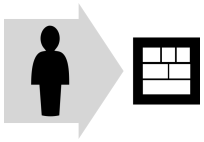
Speech recognition (Vocapia)



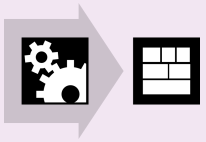
Recognition correction (manual)



Conversion to Praat TextGrid (script)

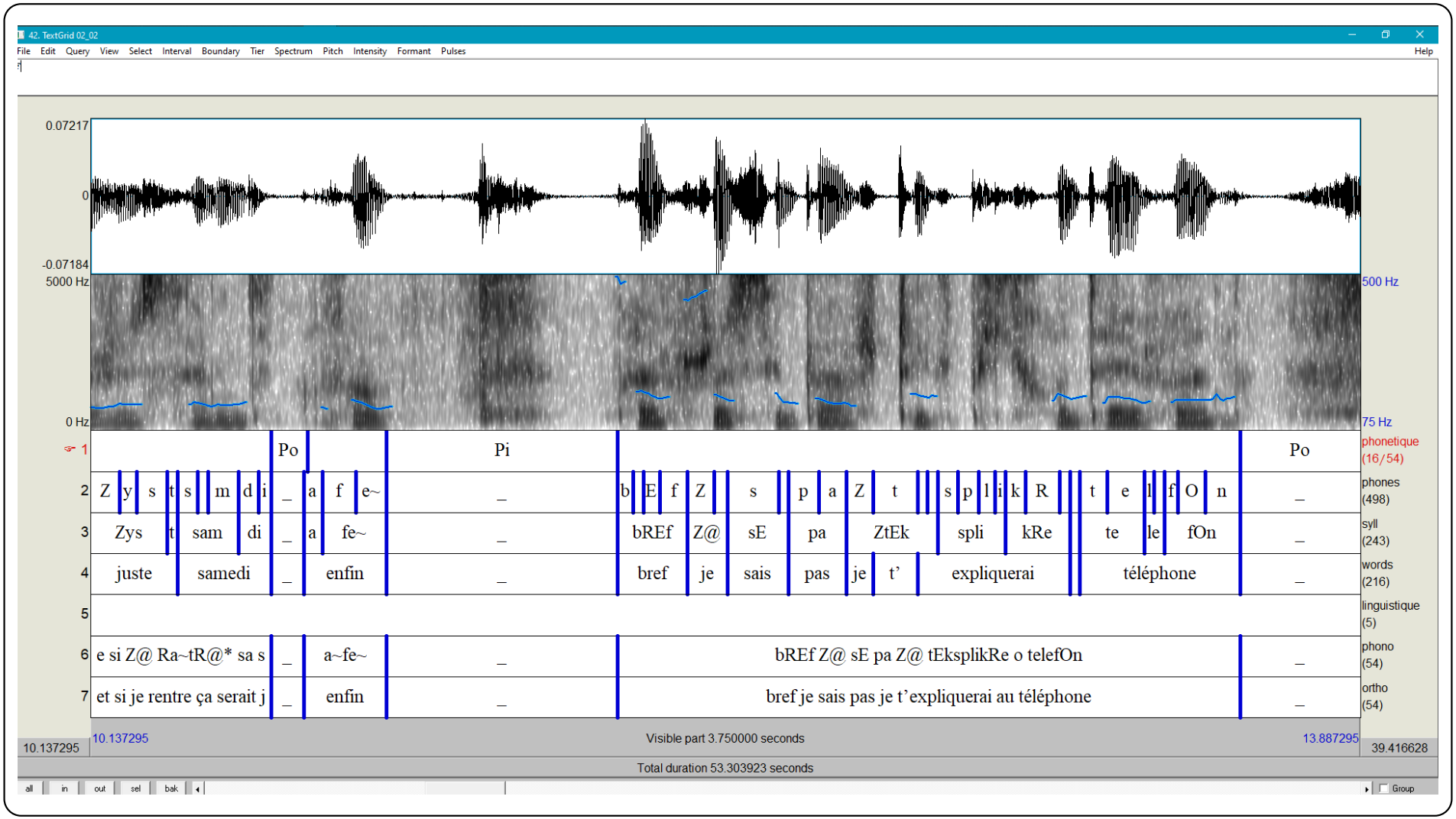


Phonetic analysis preparation (manual)



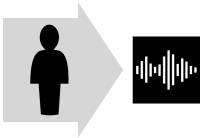
Phonetic analysis (EasyAlign)

Chaîne de traitement

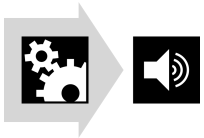


Analyse phonétique automatique (EasyAlign)

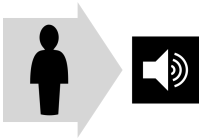
Chaîne de traitement



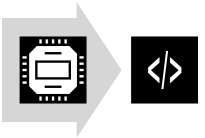
Vocal production (misc. platforms)



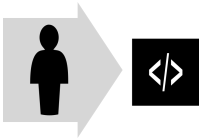
Conversion to WAV (audacity)



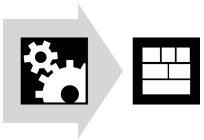
Signal anonymization (manual)



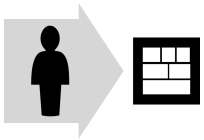
Speech recognition (Vocapia)



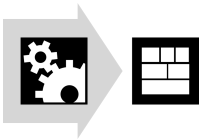
Recognition correction (manual)



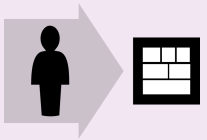
Conversion to Praat TextGrid (script)



Phonetic analysis preparation (manual)

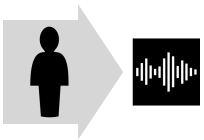


Phonetic analysis (EasyAlign)

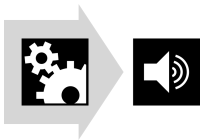


Phonetic and word correction (manual)

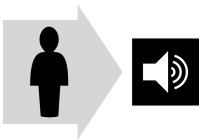
Chaîne de traitement



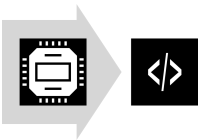
Vocal production (misc. platforms)



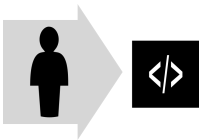
Conversion to WAV (audacity)



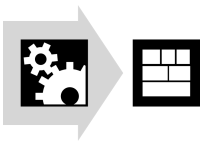
Signal anonymization (manual)



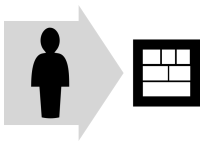
Speech recognition (Vocapia)



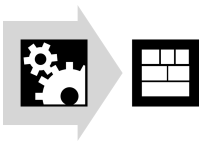
Recognition correction (manual)



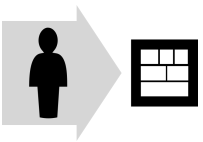
Conversion to Praat TextGrid (script)



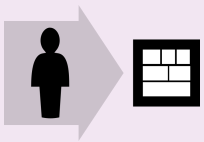
Phonetic analysis preparation (manual)



Phonetic analysis (EasyAlign)

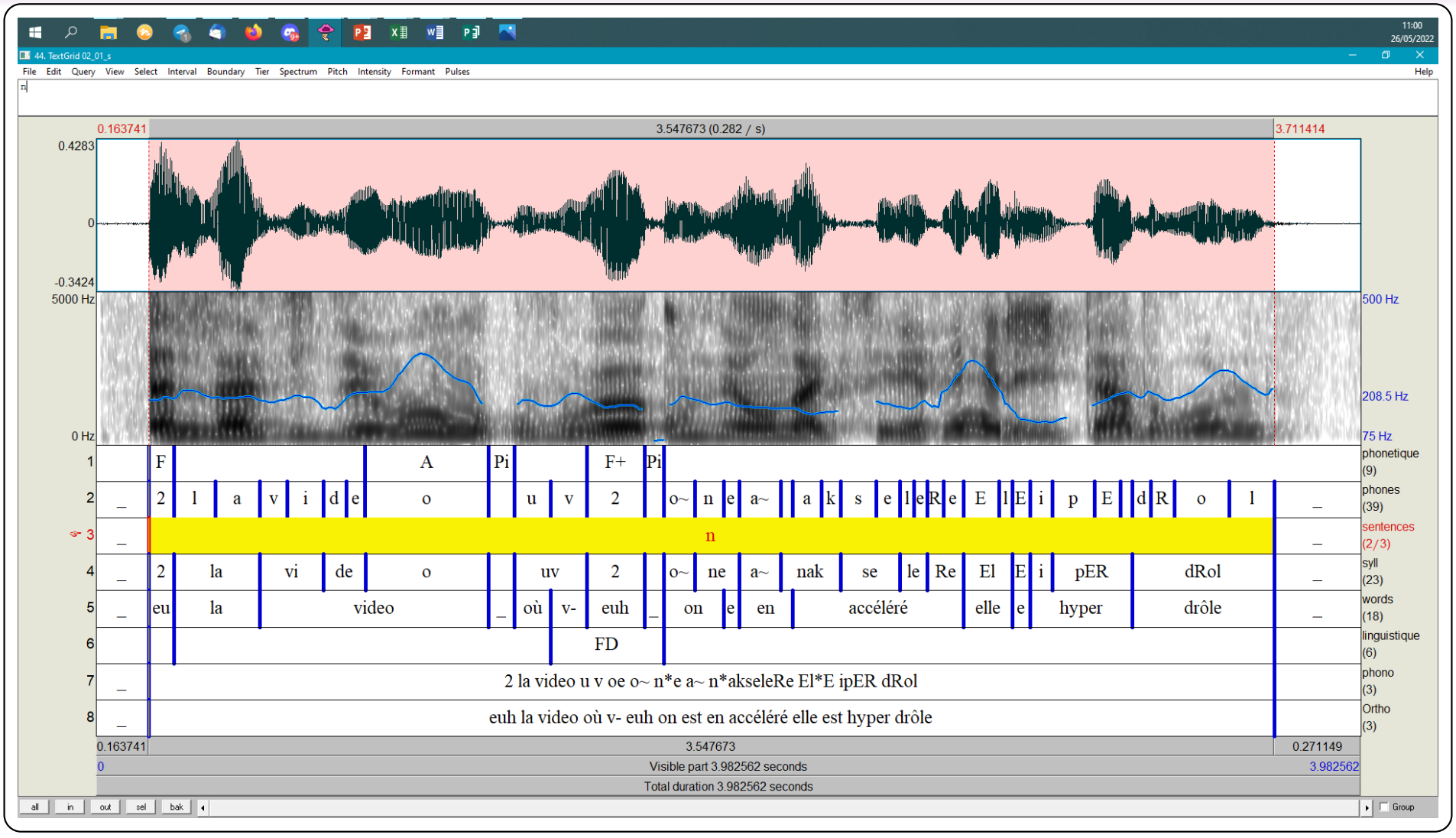


Phonetic and word correction (manual)



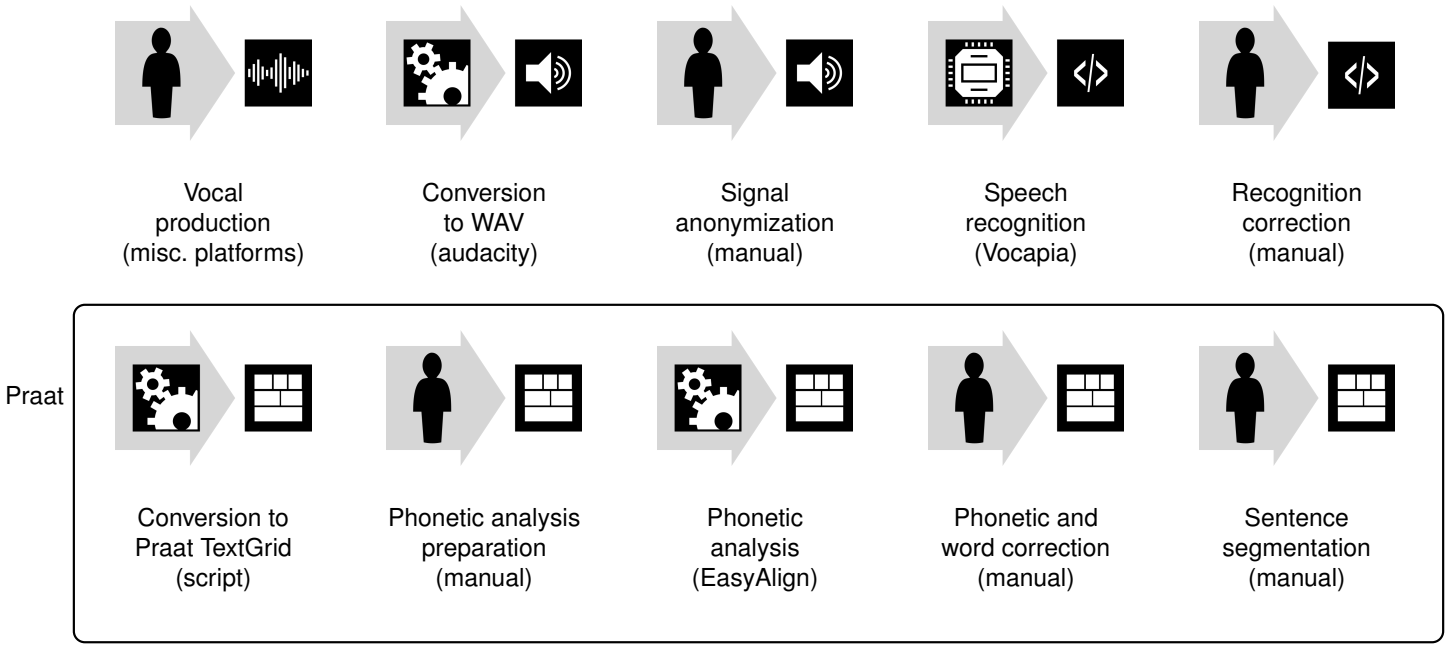
Sentence segmentation (manual)

Chaîne de traitement

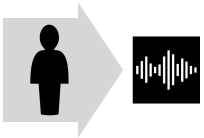


Découpage manuel en phrases

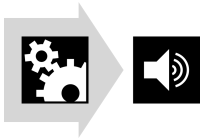
Chaîne de traitement



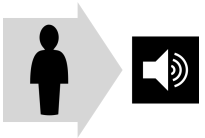
Chaîne de traitement



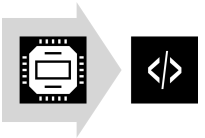
Vocal production (misc. platforms)



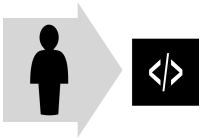
Conversion to WAV (audacity)



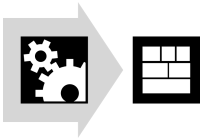
Signal anonymization (manual)



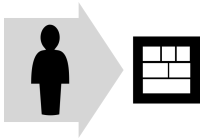
Speech recognition (Vocapia)



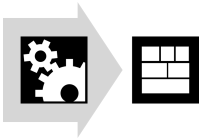
Recognition correction (manual)



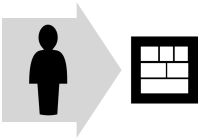
Conversion to Praat TextGrid (script)



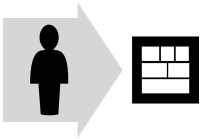
Phonetic analysis preparation (manual)



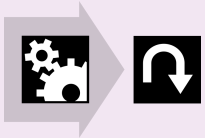
Phonetic analysis (EasyAlign)



Phonetic and word correction (manual)

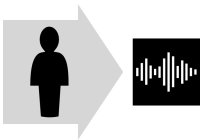


Sentence segmentation (manual)

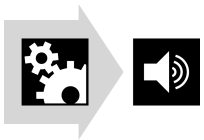


Conversion to CoNLL (script)

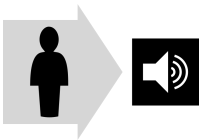
Chaîne de traitement



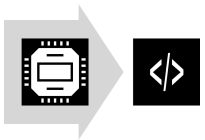
Vocal production (misc. platforms)



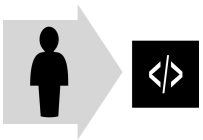
Conversion to WAV (audacity)



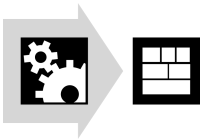
Signal anonymization (manual)



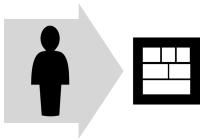
Speech recognition (Vocapia)



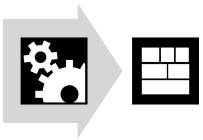
Recognition correction (manual)



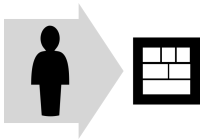
Conversion to Praat TextGrid (script)



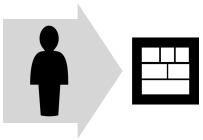
Phonetic analysis preparation (manual)



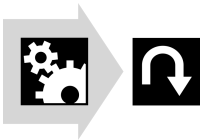
Phonetic analysis (EasyAlign)



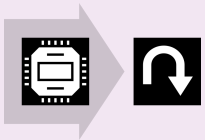
Phonetic and word correction (manual)



Sentence segmentation (manual)

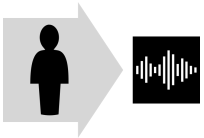


Conversion to CoNLL (script)

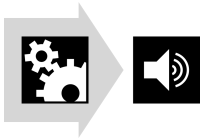


Automatic annotation (morpho, syntax, lemmata) (Trankit)

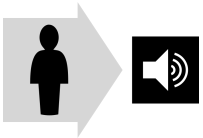
Chaîne de traitement



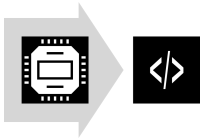
Vocal production (misc. platforms)



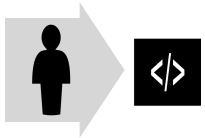
Conversion to WAV (audacity)



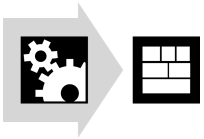
Signal anonymization (manual)



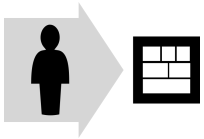
Speech recognition (Vocapia)



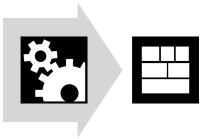
Recognition correction (manual)



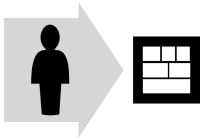
Conversion to Praat TextGrid (script)



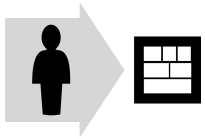
Phonetic analysis preparation (manual)



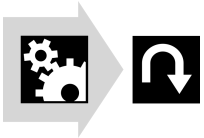
Phonetic analysis (EasyAlign)



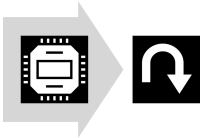
Phonetic and word correction (manual)



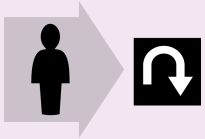
Sentence segmentation (manual)



Conversion to CoNLL (script)

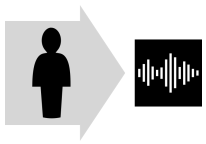


Automatic annotation (morpho, syntax, lemmata) (Trankit)

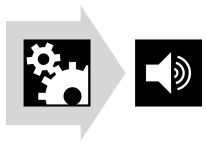


Morpho, syntax, lemmata correction (manual)

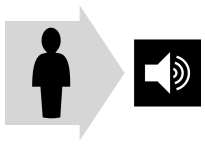
Chaîne de traitement



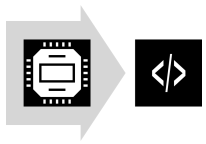
Vocal production (misc. platforms)



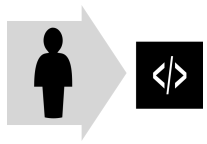
Conversion to WAV (audacity)



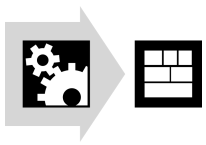
Signal anonymization (manual)



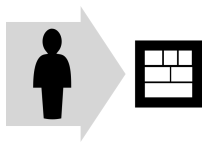
Speech recognition (Vocapia)



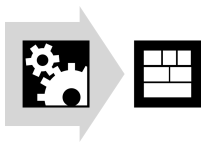
Recognition correction (manual)



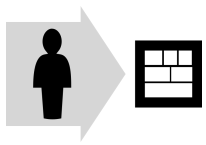
Conversion to Praat TextGrid (script)



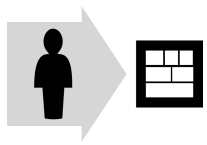
Phonetic analysis preparation (manual)



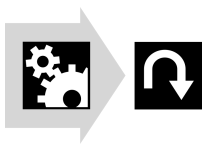
Phonetic analysis (EasyAlign)



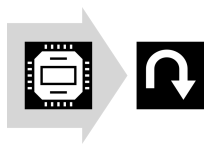
Phonetic and word correction (manual)



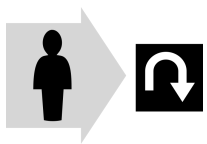
Sentence segmentation (manual)



Conversion to CoNLL (script)



Automatic annotation (morpho, syntax, lemmata) (Trankit)



Morpho, syntax, lemmata correction (manual)

Final corpus (script)

Formats et Outils

Formats et Outils

Objectifs pratiques

- Diffusion
- Extractions
- Vérifications croisées (cf. corrections manuelles)

	A	B	C	D	E
5525	... papa ... j' ai	trouvé	un portefeuille ... et tu sais au mec et je	tRu.ve	02_09
5526	ça redevenait un blob ... et genre je commençais à	trouver	ça vraiment bizarre ... et ... eh ben c' était	tRu.ve	06_07_1
5527	pas ... en soi j' aime bien l' idée du	truc	mais euh peut-être pas avec le même fonctionnement qu' on	tRuk	07_10
5528	alors je vais te sortir le	truc	bateau ... mais écoute ... vraiment ce que toi tu	tRy.k@	05_05
5529	euh réponds moi ... au ... du ... coup au	truc	euh ... pour remplir ... et après je t' appelle	tRy.k2	02_02
5530	en fait euh ... faut que je discute d' un	truc	avec toi je sais pas si je rentre pour aller	tRy.ka	02_02
5531	une grande toile ... et chacun euh doit peindre un	truc	avec son doigt et tout ... moi je trouve ça	tRy.ka	05_04
5532	comme si euh comme si le seul ... le seul	truc	... qui était ennuyeux dans cette affaire ... euh c'	tRyk	01_03
5533	je crois pas ... et euh sinon ... le vrai	truc	que faut regarder ... c' est euh ... l' orthotypographie	tRyk	02_08
5534	donc du coup j' ai pas pu faire encore le	truc	de #name mais je lui ferai euh son mug magique	tRyk	05_01
5535	bon ... ben t' as été euh coupé dans ton	truc	... et euh sinon oui euh finalement j' ai vu	tRyk	05_02
5536	ai dû chercher chercher chercher chercher c' était un petit	truc	de merde ... mais bref j' ai galéré ... et	tRyk	05_06
5537	ah t' as pas tes règles qui arrivent ou un	truc	comme ça ... moi ça me le fait quand j'	tRyk	06_03
5538	sait qu' il enfin ... que si il fait ce	truc-	là ... ça va pas bien se ... passer ça	tRy.kla	01_03
5539	non mais là ... j' ai trop de	trucs	à dire ... euh ... on en parle du DLC	tRy.ka	07_05
5540	#name ... mais vu que #name elle perd souvent les	trucs	du coup je vous l' enverrai par la poste mais	tRyk	02_02
5541	... et tu aurais tu aurais moins travaillé à tes	trucs	à toi ... ou tu aurais pas travaillé à des		01_01
5542	à toi ... ou tu aurais pas travaillé à des	trucs	à toi enfin je pense pas ... et euh ...		01_01
5543	... #indistinct euh ... que ça va bien ... que	tu	as bien travaillé au mois d' aout euh ... c'	ta	01_01
5544	fait c' est pas mal ... parce que ... si	tu	avais ... enfin tu vois ... si tu avais été	ta	01_01
5545	... si tu avais ... enfin tu vois ... si	tu	avais été prise là en fait ... bien précisément ...	ta	01_01
5546	tendu ... mais en même temps je ... pense que	tu	as assez ... enfin tu as suffisamment d' expérience ...	ta	01_01
5547	temps je ... pense que tu as assez ... enfin	tu	as suffisamment d' expérience ... et de sens de l'	ta	01_01
5548	et tout ... et euh ... et le travail que	tu	as fait en aout ben tu l' as fait quoi	ta	01_01
5549	et euh sinon oui euh finalement j' ai vu que	tu	avais écrit #name à côté de ... enfin dans le	ta	05_02
5550	... annulé ... non c' est pas le mot mais	tu	as compris ... fermé quoi ... euh je sais pas	ta	07_10
5551	et puis sinon je te donne l' info euh ...	tu	en fais ce que tu veux voilà ...	ta~	07_04
5552	toi tu fais quand même gaffe ... autant que si	tu	étais pas fat- euh vacciné ... tu tu t' en	te	04_02
5553	hein c' est tout ... genre euh c' est m-	tu	es en aucun cas dans un dans le problème là	te	04_04
5554	... bon ben tant pis ça y est	tu	fais la gueule ou quoi ...	tfE	06_08_2
5555	prise là en fait ... bien précisément ... en fait	tu	y aurais passé ton mois d' aout ... et euh	ti	01_01
5556	passé ton mois d' aout ... et euh ... et	tu	aurais tu aurais moins travaillé à tes trucs à toi	to	01_01
5557	même euh ... juste en lui parlant je pense que	tu	aurais pas besoin de lui dire grand chose parce que	to	05_05
5558	pas quoi lui dire ... mais je suis sûre que	tu	aurais même pas besoin de lui dire grand- chose ...	to	05_05
5559	j' avais acheté un blob ... je sais pas si	tu	vois ce que c' est c' est ... les petites	tu	06_07_1
5560	... euh ... moi j' ai l' impression enfin	tu	vois que ce genre ... d' acte manqué ... en	ty	01_01
5561	mal ... parce que ... si tu avais ... enfin	tu	vois ... si tu avais été prise là en fait	ty	01_01
5562	et le travail que tu as fait en aout ben	tu	l' as fait quoi ... euh donc euh moi j'	ty	01_01
5563	surtout chez les enfants euh ... euh ... un peu	tu	vois qui sont plus pas des tous petits ... qui	ty	01_03
5564	leur qu' on leur donne ... mais là ... enfin	tu	vois #name qui répète je suis désolé je suis désolé	ty	01_03
5565	désolé je suis désolé ... mais en fait ... enfin	tu	vois c' est comme euh quand tu dis parfois euh	ty	01_03
5566	fait ... enfin tu vois c' est comme euh quand	tu	dis parfois euh #name mais en fait euh ... on	ty	01_03
5567	mais en fait euh ... on s' en fout que	tu	sois désolé ... parce que moi si je te demande	ty	01_03
5568	est utile pour toi ... donc euh ... enfin ...	tu	vois comme si euh comme si le seul ... le	ty	01_03

Vérification de la segmentation syllabique

Formats et Outils

Objectifs pratiques

- Diffusion
- Extractions
- Vérifications croisées (cf. corrections manuelles)

Sélection des formats

Une diffusion « fair » :

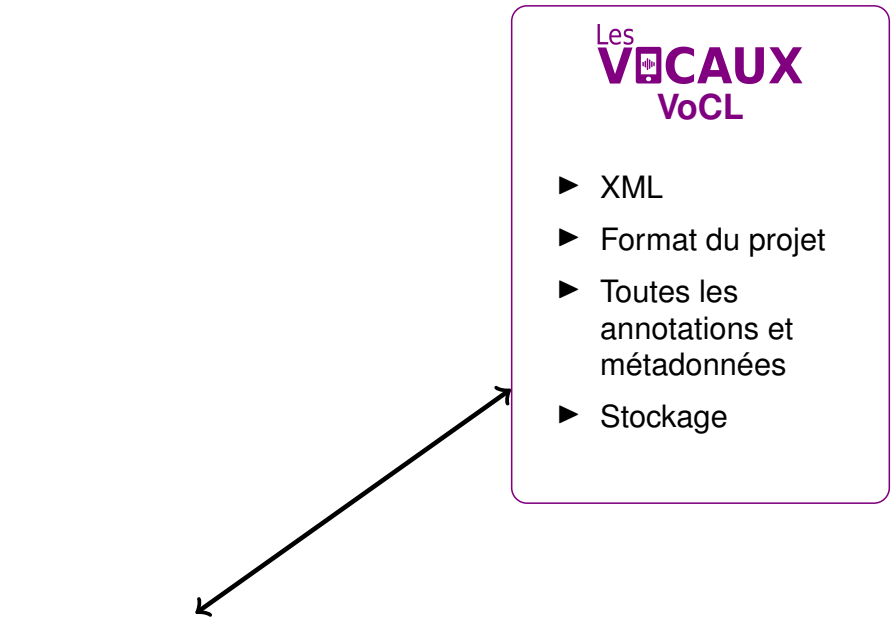
- Formats ouverts
- Formats bien connus
- Formats adaptés aux données textuelles

Formats et Outils

Les VOCAUX
VoCL

- XML
- Format du projet
- Toutes les annotations et métadonnées
- Stockage

Formats et Outils



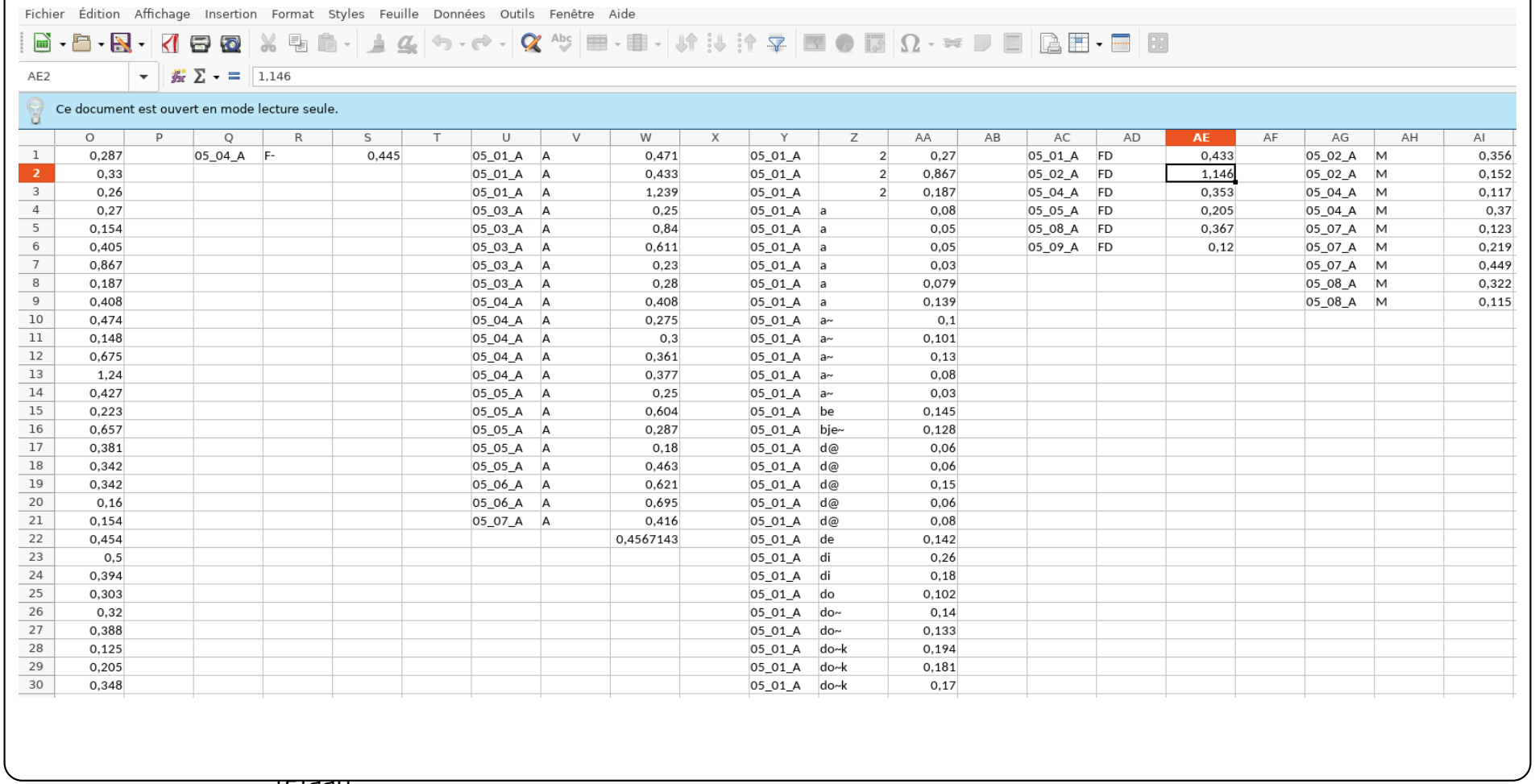
Les VOCAUX
VoCL

- XML
- Format du projet
- Toutes les annotations et métadonnées
- Stockage

TextGrid

- Spécifique à Praat
- Toutes les annotations
- Annotation et analyse du signal
- Requêtes « unidimensionnelles » (Praat)

Formats et Outils

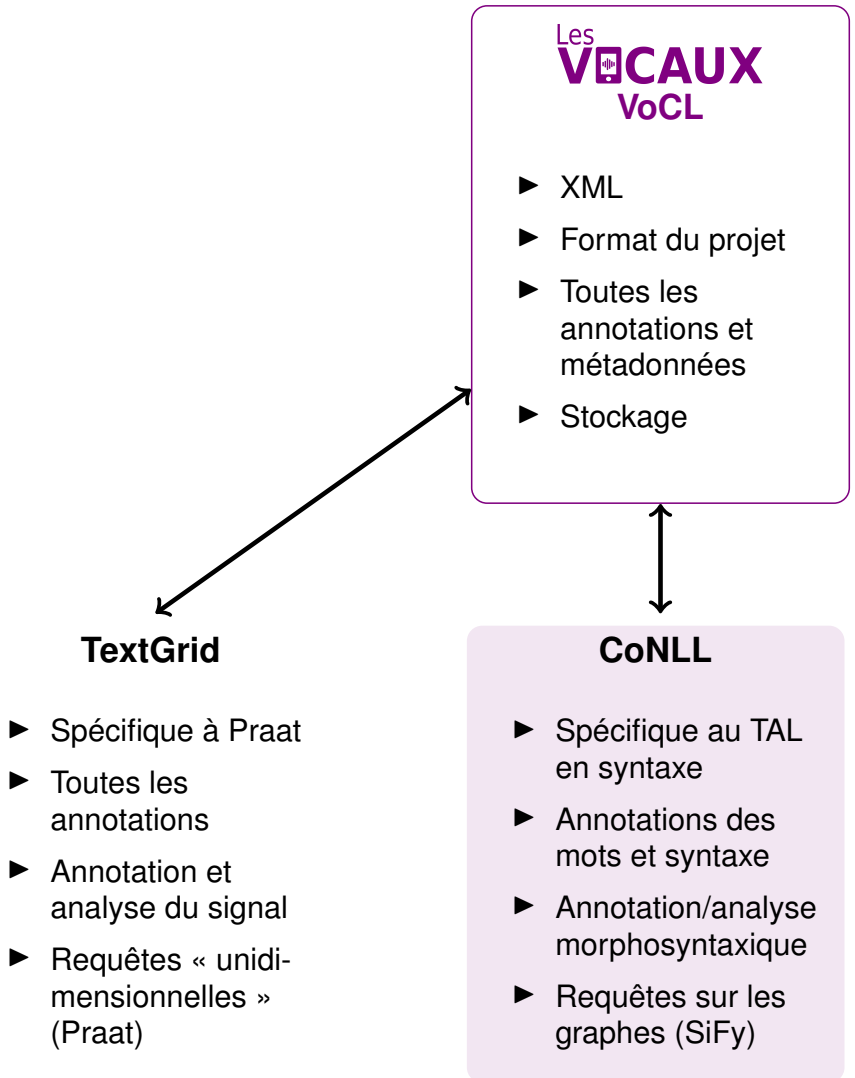


	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	0,287		05_04_A	F-	0,445		05_01_A	A	0,471		05_01_A	2	0,27		05_01_A	FD	0,433		05_02_A	M	0,356
2	0,33						05_01_A	A	0,433		05_01_A	2	0,867		05_02_A	FD	1,146		05_02_A	M	0,152
3	0,26						05_01_A	A	1,239		05_01_A	2	0,187		05_04_A	FD	0,353		05_04_A	M	0,117
4	0,27						05_03_A	A	0,25		05_01_A	a	0,08		05_05_A	FD	0,205		05_04_A	M	0,37
5	0,154						05_03_A	A	0,84		05_01_A	a	0,05		05_08_A	FD	0,367		05_07_A	M	0,123
6	0,405						05_03_A	A	0,611		05_01_A	a	0,05		05_09_A	FD	0,12		05_07_A	M	0,219
7	0,867						05_03_A	A	0,23		05_01_A	a	0,03						05_07_A	M	0,449
8	0,187						05_03_A	A	0,28		05_01_A	a	0,079						05_08_A	M	0,322
9	0,408						05_04_A	A	0,408		05_01_A	a	0,139						05_08_A	M	0,115
10	0,474						05_04_A	A	0,275		05_01_A	a~	0,1								
11	0,148						05_04_A	A	0,3		05_01_A	a~	0,101								
12	0,675						05_04_A	A	0,361		05_01_A	a~	0,13								
13	1,24						05_04_A	A	0,377		05_01_A	a~	0,08								
14	0,427						05_05_A	A	0,25		05_01_A	a~	0,03								
15	0,223						05_05_A	A	0,604		05_01_A	be	0,145								
16	0,657						05_05_A	A	0,287		05_01_A	bje~	0,128								
17	0,381						05_05_A	A	0,18		05_01_A	d@	0,06								
18	0,342						05_05_A	A	0,463		05_01_A	d@	0,06								
19	0,342						05_06_A	A	0,621		05_01_A	d@	0,15								
20	0,16						05_06_A	A	0,695		05_01_A	d@	0,06								
21	0,154						05_07_A	A	0,416		05_01_A	d@	0,08								
22	0,454								0,4567143		05_01_A	de	0,142								
23	0,5										05_01_A	di	0,26								
24	0,394										05_01_A	di	0,18								
25	0,303										05_01_A	do	0,102								
26	0,32										05_01_A	do~	0,14								
27	0,388										05_01_A	do~	0,133								
28	0,125										05_01_A	do~k	0,194								
29	0,205										05_01_A	do~k	0,181								
30	0,348										05_01_A	do~k	0,17								

(Praat)

Décomptes Praat

Formats et Outils



localhost/~mzi/syfi/

Menu Open paste basket Get current state

SPARQL Query | GREW Query

```
% In order to build the concordance and to sort results,
% a node named "FOC" must be defined. Nodes numbered C1, C2
% C3, etc. are highlighted in the treeview

pattern {
    V -[nsubj]-> FOC;
}
```

Powered by Bruno Guillaume's Grew

Sort keys

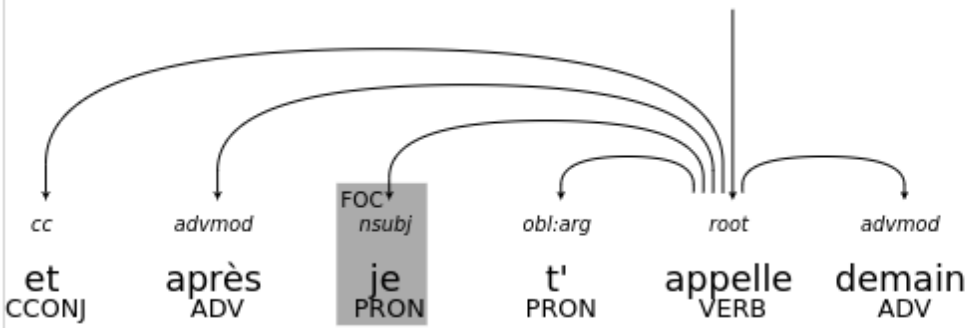
Ref. node	Sort key	Actions
FOC	Lemma	
FOC	UPOS	

Options

Output format: On screen

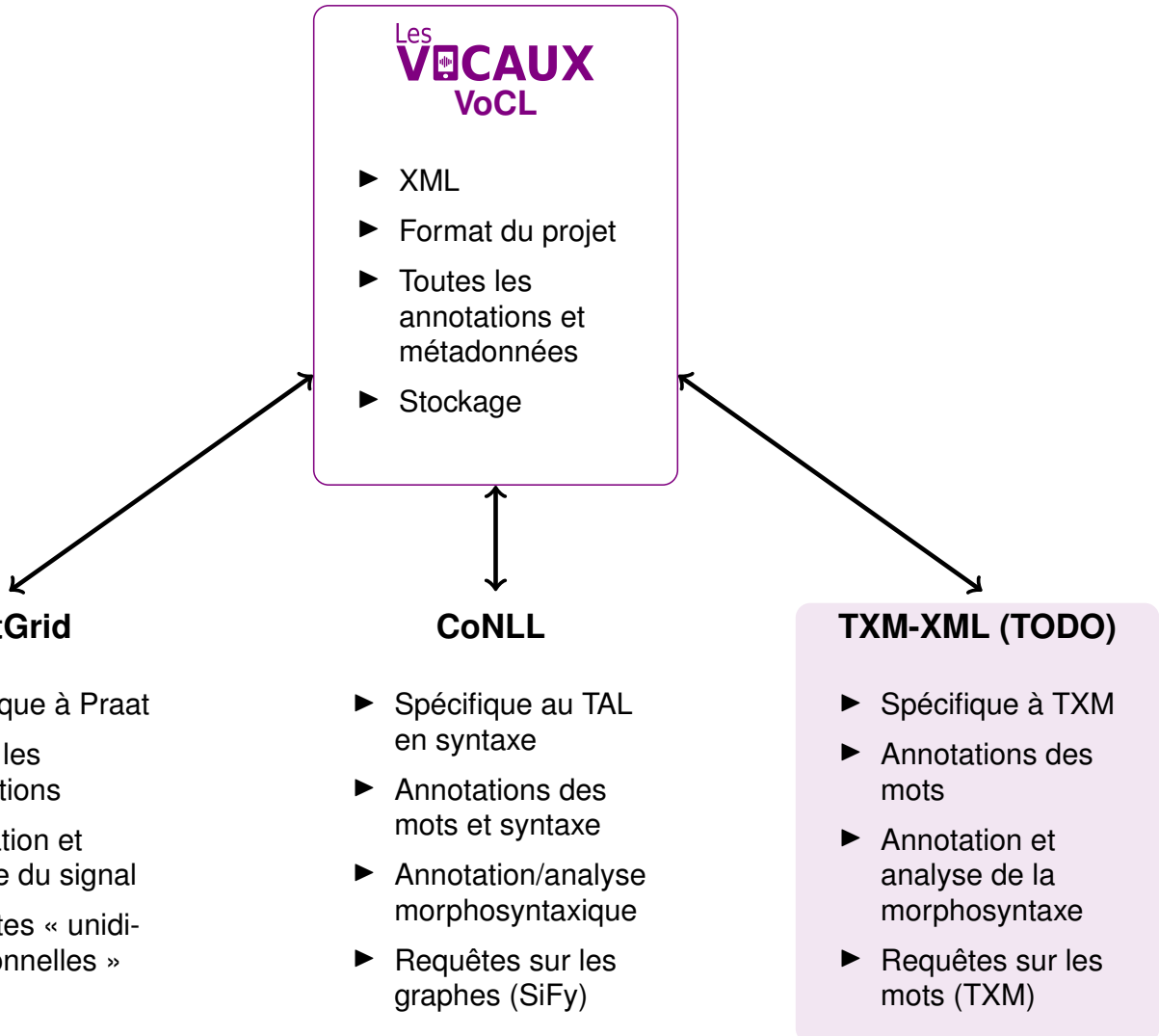
Results : 24 hits

N	FOC.lemma	FOC.upos					
1	il	PRON	et après	**je**	t' appelle demain		LesVocaux
2	il	PRON	mais ça	**tu**	me rediras		LesVocaux
3	il	PRON	mais vu que #name	**elle**	perd souvent les trucs du coup je vous l' enverrai		LesVocaux

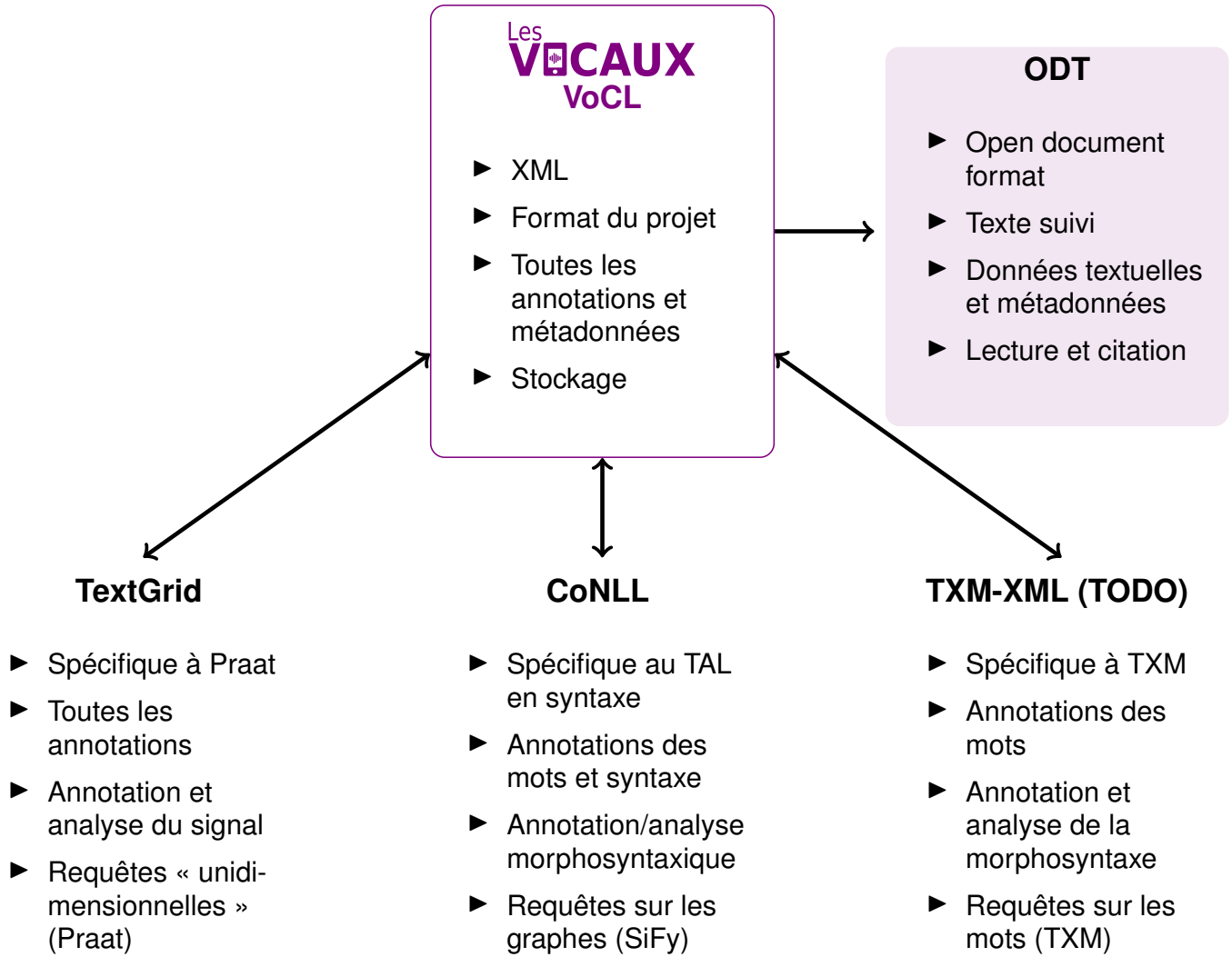


Requête dans Syfi

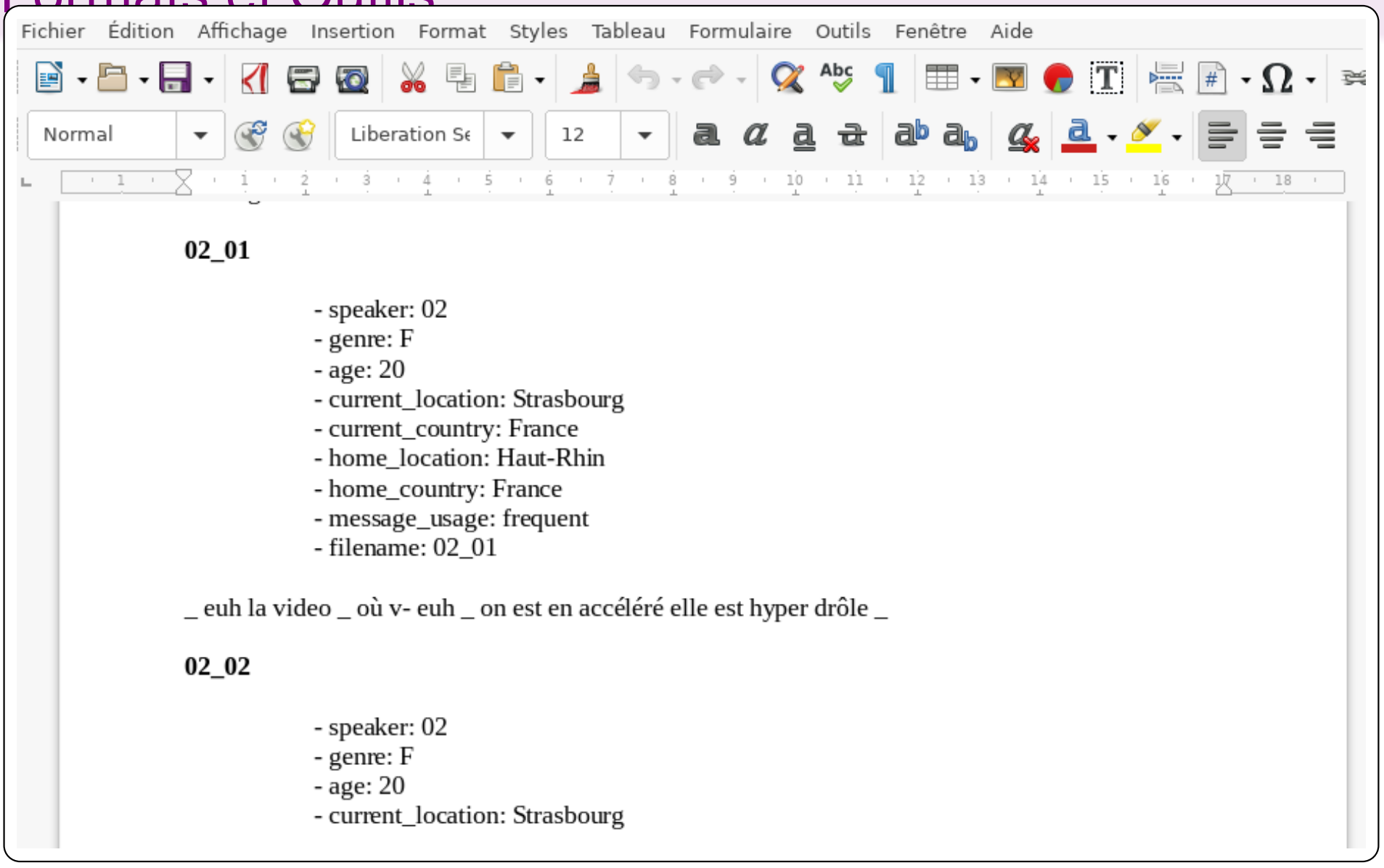
Formats et Outils



Formats et Outils



Formats et Outils



Lecture « simple »

Conclusion

Conclusion

Nous avons présenté :

Conclusion

Nous avons présenté :

- Intérêt/type des données

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil
- Phénomènes saillants

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil
- Phénomènes saillants
- Chaîne de traitement (beaucoup d'étapes, mixage manuel/automatique)

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil
- Phénomènes saillants
- Chaîne de traitement (beaucoup d'étapes, mixage manuel/automatique)
- Format variable des sorties (permet l'interopérabilité)

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil
- Phénomènes saillants
- Chaîne de traitement (beaucoup d'étapes, mixage manuel/automatique)
- Format variable des sorties (permet l'interopérabilité)

Pistes d'exploitation

- Variation diatopique (origine des sujets parlants)
- Variation diaphasique
- Microdiachronie (structures émergentes, modes)

Conclusion

Nous avons présenté :

- Intérêt/type des données
- Méthodologie de recueil
- Phénomènes saillants
- Chaîne de traitement (beaucoup d'étapes, mixage manuel/automatique)
- Format variable des sorties (permet l'interopérabilité)

Pistes d'exploitation

- Variation diatopique (origine des sujets parlants)
- Variation diaphasique
- Microdiachronie (structures émergentes, modes)

<glikman@unistra.fr>
<cfauth@unistra.fr>
<nicolas.mazziotta@uliege.be>
<christophe.benzitoun@univ-lorraine.fr>