Discours VALIBEL Variation

La base de données textuelles orales **VALIBEL**

de la variation aux variétés

http://www.uclouvain.be/valibel.html

Alice Bardiaux (FNRS – UCL)

parcours :
du corpus à l'étude de la variation

- centre de recherche VALIBEL
- base de données textuelles orales VALIBEL
  - bref historique : un corpus pour l'étude de la variation
  - philosophie et principes généraux
  - composition du corpus et métadonnées
  - transcription et annotation des données
  - interface de consultation MOCA
- au-delà du corpus : réflexion sur la notion de *variation* et *variété*

le centre de recherche VALIBEL

VALIBEL (1989)

VAriétés **LI**nguistiques

du français en Belgique

M. Francard (dir.)

A.C. Simon, Ph. Hambye

CETIS

Centre d'**E**tudes

du **T**exte et di d**IS**cours

L. Degand (dir.)

création de la « banque de données textuelles orales Valibel » (1989)

Discours VALIBEL Variation

création en 2009 (A. C. Simon, dir.)

le centre de recherche VALIBEL

aujourd'hui…

domaines

- analyse du discours
- sociolinguistique
- prosodie
- linguistique de corpus

niveaux linguistiques

- lexicologie
- phonétique, phonologie
- sémantique
- pragmatique

données langagières **authentiques**
prises dans leur **contexte**
de production effective

pratiques
vs.
normes
& représentations

linguistique théorique
vs.
linguistique appliquée

base de données VALIBEL

naissance en **1989**

influence de **2** courants

GARS
(Blanche-Benveniste, JeanJean)

- priorité aux phénomènes morphologiques et syntaxiques

sociolinguistes québécois
(Cedergren, Sankoff, Deshaies…)

- intégration de la récolte et de la transcription de corpus oraux
- démarche variationniste
- données sociolinguistique et situationnelles

base de données VALIBEL

objectifs et philosophie = documenter la diversité des pratiques

- étude de la variation > métadonnées
- usages attestés et fiabilité des données
- numérisation, transcription, alignement son-texte
- données belges
  > concertation avec des équipes étrangères
  > intégration dans un approche globale : BDLP, PFC

http://www.tlfq.ulaval.ca/bdlp/ et www.projet-pfc.net

alimentation de la base de données

- apports des chercheurs « juniors » : mémorants, doctorants
  > thématiques de recherche variées
  > corpus encodés variés

→nécessité de protocoles précis

base de données VALIBEL

en bref…

- 45 corpus (1987 – 2012)
- 1000 enregistrements
- 729 informateurs → de Bruxelles et de Wallonie
- +/- 5 millions de mots
- diversité des genres :
  conversations informelles, entretiens (semi)guidés, débats, journaux radiophoniques et télévisés, interviews politiques et culturelles, discours académique…

→contribution au projet ORFEO (projet ANR)
« outils et ressources pour le français écrit et oral »
http://www.lattice.cnrs.fr/ORFEO-Outils-et-Ressources-pour-le

base de données VALIBEL

en bref…

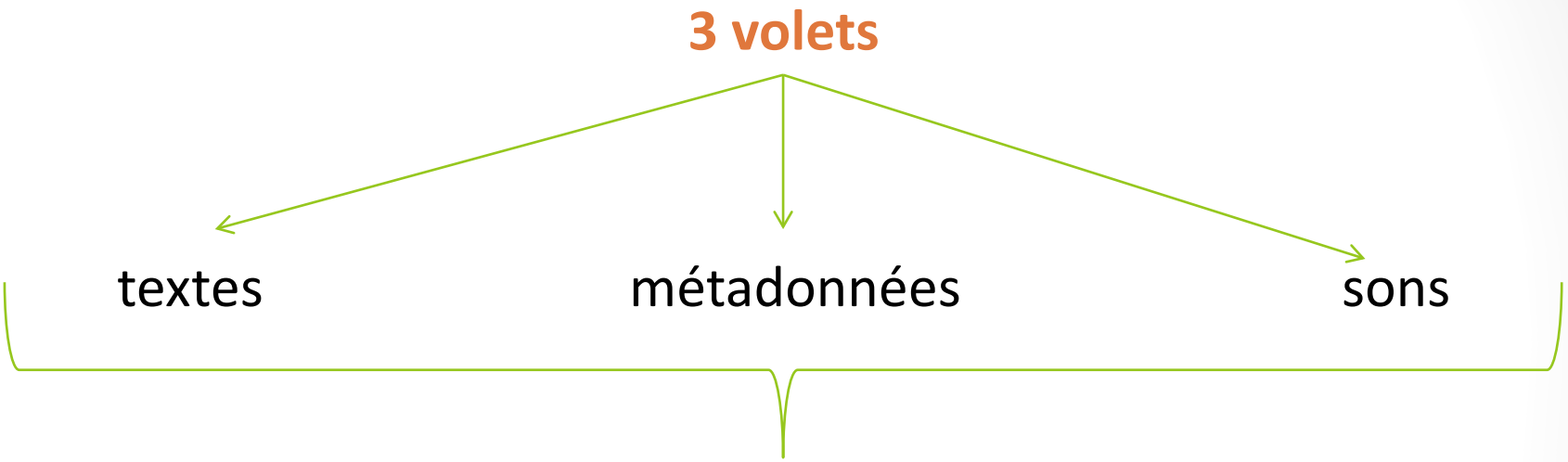
- 45 corpus (1987 – 2012) → multiples sources vs. ~~corpus clos~~
- 1000 enregistrements
- 729 informateurs → de Bruxelles et de Wallonie
- +/- 5 millions de mots
- diversité des genres :
  conversations informelles, entretiens (semi)guidés, débats, journaux radiophoniques et télévisés, interviews politiques et culturelles, discours académique…

base de données VALIBEL

45 corpus – 5 axes de recherche

- sociolinguistique > insécurité linguistique
  - corpus de thèses (cf. Hambye, Audrit)
  - corpus de mémoires
- phonétique, phonologie, prosodie
  - projet international Phonologie du Français Contemporain (PFC) http://www.projet-pfc.net/
  - corpus de thèses (cf. Hambye, Audrit, Bardiaux…) et de mémoires
- lexicologie différentielle et lexicographie
  - dictionnaire des belgicismes (cf. Francard)
- analyse du discours > marqueurs de discours et de disfluence (cf. Degand)
  - réseau européen COST TextLink : structuring discourse in multilingual Europe
  - projet ARC : Fluency and disfluency markers
- vieillissement langagier > corpus Corpage (cf. Bolly)

base de données VALIBEL



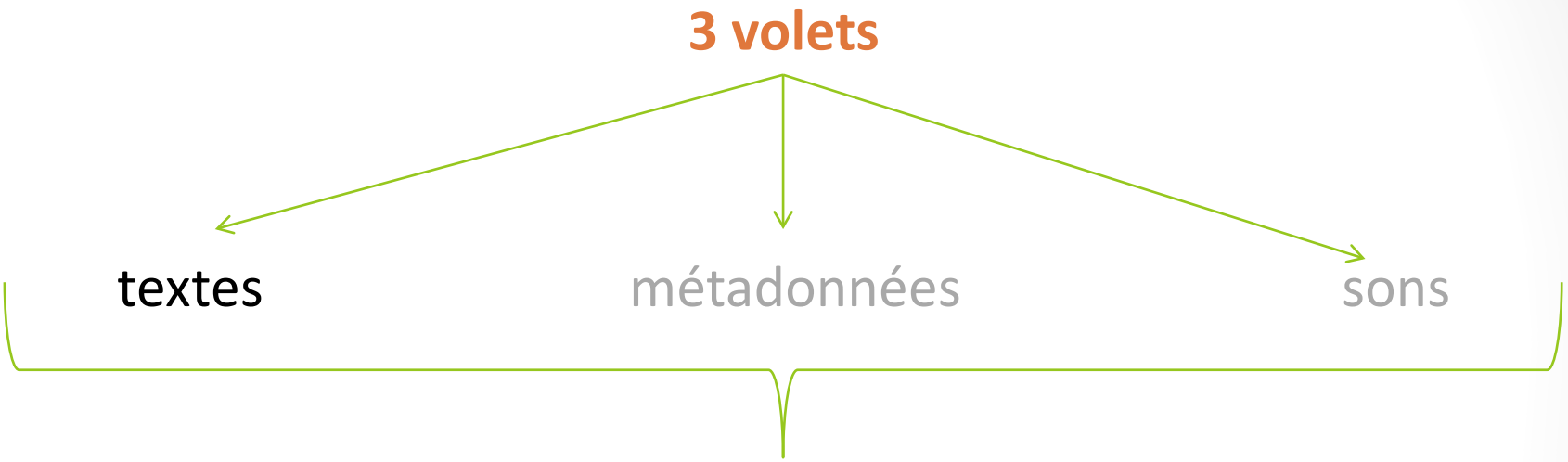
interface de consultation **MOCA** (sous firefox)
http://moca.fltr.ucl.ac.be/moca/index.php

75% des corpus = diffusables et libres de droit

- 5% transcrits et alignés en tours de parole
- 70% transcrits

NB : base de données en cours de mise à jour
(harmonisation, anonymisation, transcription, alignement, encodage des métadonnées)

base de données VALIBEL



interface de consultation **MOCA** (sous firefox)

base de données VALIBEL

TEXTE > transcriptions > **conventions de transcription**

principes généraux

- orthographe standard > pas de trucage orthographique
- prise en compte de l'oralité du corpus
  - pauses, silences
  - répétitions, corrections, reprises et mots tronqués
    > marques du travail de formulation
  - phénomènes liés à l'interaction
    > tours de parole et chevauchements
- compatibilité avec un traitement informatisé des données

base de données VALIBEL

TEXTE > transcriptions > **conventions de transcription**

principes généraux

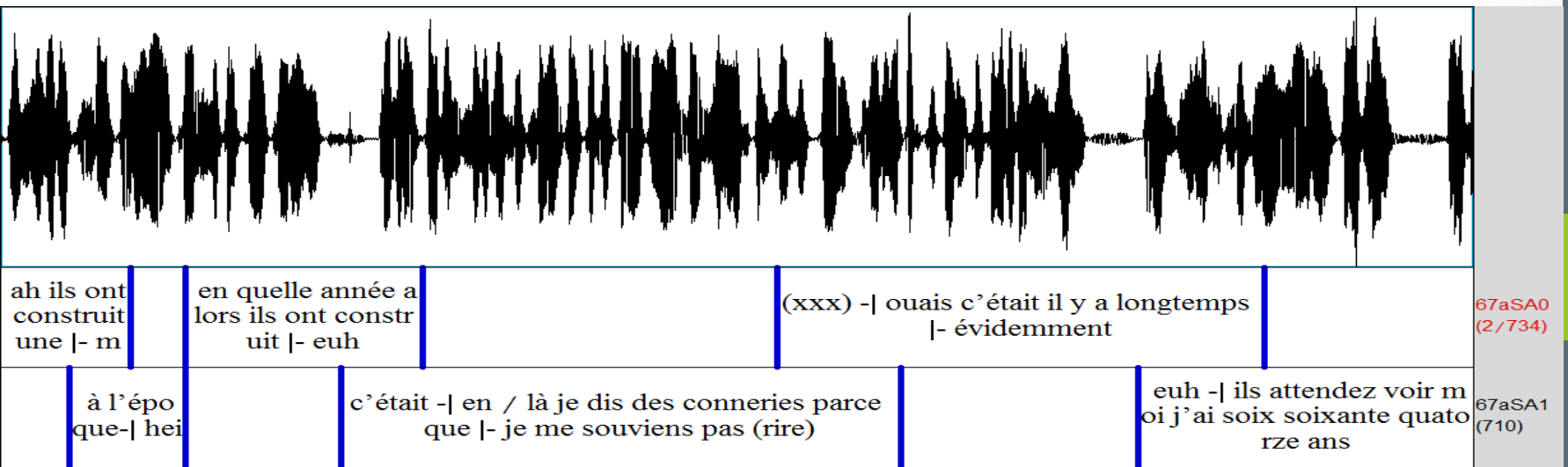
- transcription orthographique vs. ~~phonétique, intonation~~
  > accès aux données orales
  > meilleure standardisation de la transcription
- transcription + vérification systématique
- conventions appliquées à l'ensemble de la base
  > transcriptions et annotations additionnelles possibles selon les exploitations particulières
  > codage supplémentaire = dans une autre version du texte

base de données VALIBEL

TEXTE > transcriptions > **conventions de transcription**

format = texte (.txt) + alignement (.TextGrid, PRAAT)

souVC1	il y a pas beau/ il y a pas b/ beaucoup de magasins
souKC1	ah il y a un terrain de foot pour les enfants à côté / une grande plaine
souGH1	ouais / tu lâches les enfants là et tu
souVC1	ah ouais il y a une grande plaine / c'est vrai
souPH1	tu pars tranquille
souGH1	(rire) (silence)
souKC1	ah mais
souGH1	on doit / on doit être beau il a dit Marc / parce que c'est la photo officielle



base de données VALIBEL

TEXTE > transcriptions > **conventions de transcription**

quelques conventions particulières

(phénomènes liés à l'interaction)

- pauses
  - brèves /
  - longues //
  - silence (silence)
  - pauses pleines *euh*
- chevauchements : |- -|
- interruption/ (sans espace)
- aspects paraverbaux et didascalies : entre parenthèses

base de données VALIBEL

Conversation – Corpus «Variation stylistique »

(silence)
stySC1 les livres c'est triste parce que franchement euh / moi ça ne m'intéresse pas // tu as tout sur internet maintenant
(silence)
styDM1 mais / il y a un livre sur BDV
stySM1 il y a rien à faire tu es vraiment internet à mort hein toi
stySC1 mais non pas à |- mort mais bon euh < styDM1 > zoum zoum zoum zoum -| tu trouves tout là-dessus c'est vrai
styDM1 oui mais dans un |- livre
stySC1 maintenant -| je ne dis plus je vais voir sur le dictionnaire je vais voir sur internet / c'est vrai
styDM1 oui mais le temps |- {que tu mettes, de tout mettre}
stySC1 il y -| a des dicos il y a cinquante-mille dicos il y a le Collins
(silence)
sur euh / sur internet tu as tout ce que tu veux
styDM1 oui mais le temps d'allumer ton ordinateur
stySC1 ah non mais |- hé c'est
stySM1 ah mais non -| elle est en permanence sur son PC hein
stySC1 non / pas à la maison hein
styDM1 (bruit) bonjour Alice (rire)

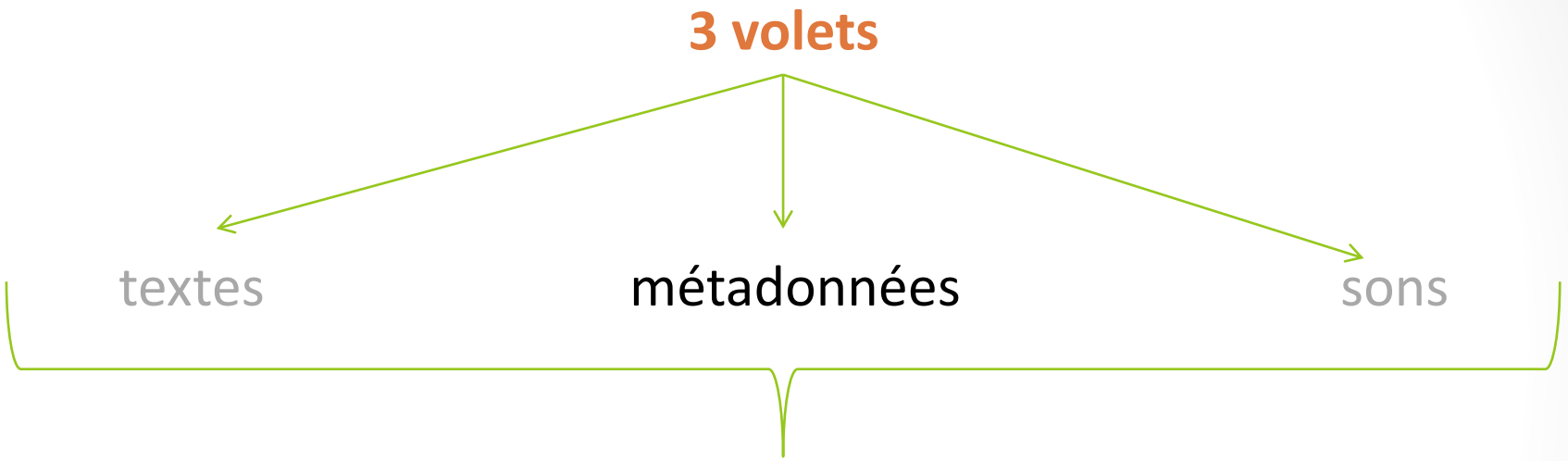
pauses (brève, longue) et silence

chevauchements

code locuteur

didascalies, commentaires

base de données VALIBEL



interface de consultation **MOCA** (sous firefox)

base de données VALIBEL

METADONNEES

- corpus
- enregistrement
- locuteur

base de données VALIBEL

METADONNEES (corpus)

Détails: corpus Accent

Masquer les détails

Modifier | Effacer

Code	acc
Date de constitution	1988
thèmes	accent régional
Description	Représentations linguistiques sur l'accent régional à Charleroi Ancien nom: CORPUS CORDIER Thérèse
enregistrements	49 (Afficher la liste)
locuteurs	53 (Afficher la liste)
Nombre de mots	16987
Durée (minutes)	
Rem. date recueil	juillet à septembre 1988
Rem. durée	39h.30
Objectif de la recherche	attitudes des locuteurs belges francophones de Charleroi et de sa région face à l'accent régional
Ancien nom	

base de données VALIBEL

METADONNEES (enregistrement)

Détails: **enregistrement stySC1s**

Masquer les détails

Modifier | Effacer | UPDATE_TRANSCRIPT

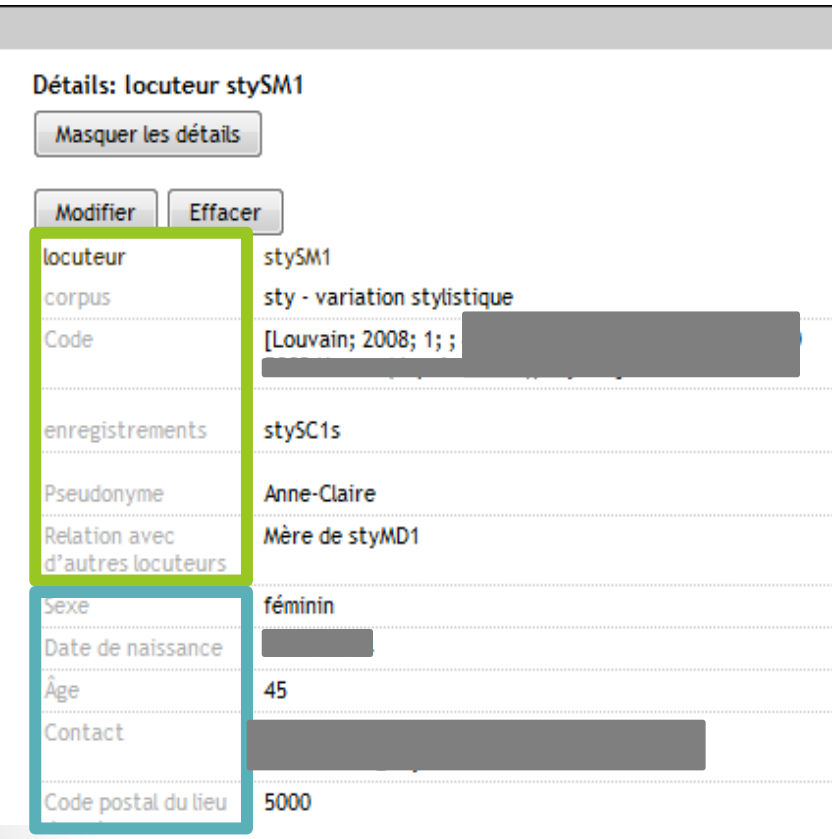
enregistrement	stySC1s (776)
corpus	sty - variation stylistique
LINK	[Louvain; 2008; stySC1s]
Documents	DOC_MAN stySC1s (son) stySC1s.wav EXPORT_TEXTGRID (Last Version)
locuteurs	stySC1 styDM1 styMA1 stySM1
Nombre de locuteurs	4 (3)
Relation entre les participants	2 soeurs 1 fils

Auteurs	[redacted]
Date de l'enregistrement	2008-03-08
Lieu de l'enregistrement	Namur (St-Gervais)
Contexte situationnel	3 personnes de la même famille autour de la table de la cuisine.
Langues	français
Type d'interaction	conversation
Durée (minutes)	
Nombre de mots	6319
Qualité du son	inconnu
Support (original)	minidisc
Support (copies)	disque dur
Satut de la transcription	non vérifié
Anonymisation	inconnu
Droits d'exploitation	inconnu
Commentaires	

nombre de locuteurs, relation entre eux, date et lieu d'enregistrement, langue, type d'interaction, durée (en min), nombre de mots, statut de l'enregistrement

base de données VALIBEL

METADONNEES (locuteur)

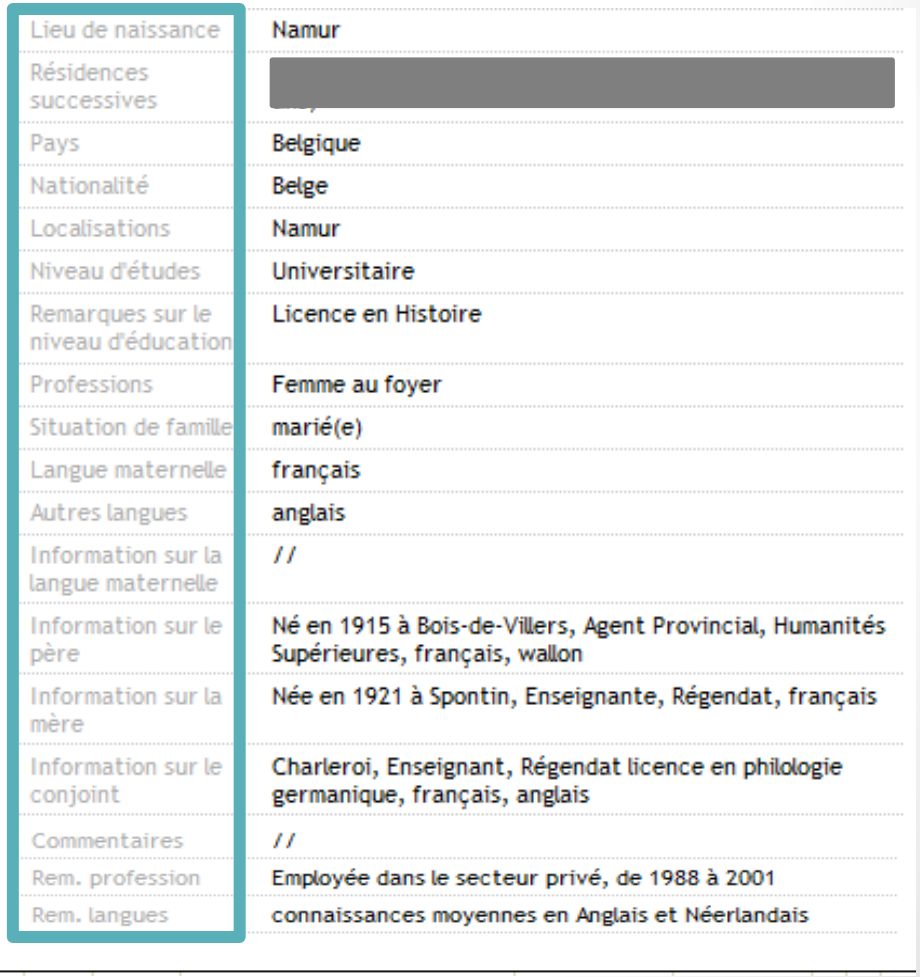


Détails: locuteur stySM1

Masquer les détails

Modifier | Effacer

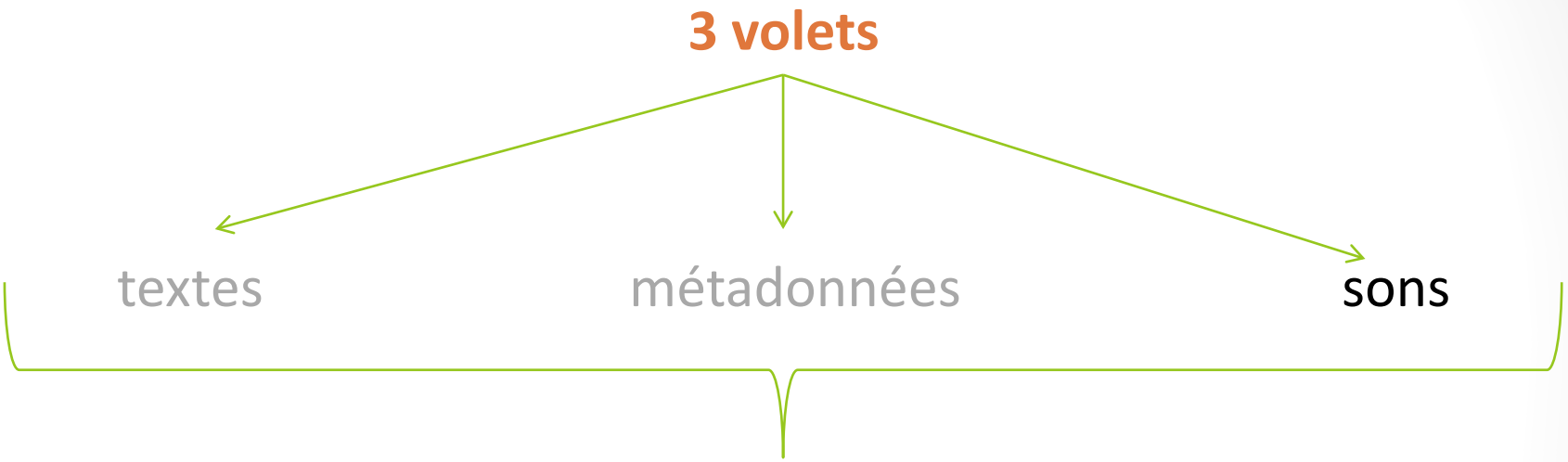
locuteur	stySM1
corpus	sty - variation stylistique
Code	[Louvain; 2008; 1; ; [redacted]
enregistrements	stySC1s
Pseudonyme	Anne-Claire
Relation avec d'autres locuteurs	Mère de styMD1
Sexe	féminin
Date de naissance	[redacted]
Âge	45
Contact	[redacted]
Code postal du lieu	5000



Lieu de naissance	Namur
Résidences successives	[redacted]
Pays	Belgique
Nationalité	Belge
Localisations	Namur
Niveau d'études	Universitaire
Remarques sur le niveau d'éducation	Licence en Histoire
Professions	Femme au foyer
Situation de famille	marié(e)
Langue maternelle	français
Autres langues	anglais
Information sur la langue maternelle	//
Information sur le père	Né en 1915 à Bois-de-Villers, Agent Provincial, Humanités Supérieures, français, wallon
Information sur la mère	Née en 1921 à Spontin, Enseignante, Régendat, français
Information sur le conjoint	Charleroi, Enseignant, Régendat licence en philologie germanique, français, anglais
Commentaires	//
Rem. profession	Employée dans le secteur privé, de 1988 à 2001
Rem. langues	connaissances moyennes en Anglais et Néerlandais

situer le locuteur au sein du corpus
+ informations personnelles

base de données VALIBEL



interface de consultation **MOCA** (sous firefox)

base de données VALIBEL

écoute d'un son

Métadonnées		Transcription					
enregistrement	locuteur	Numéro de ligne	Transcription	Temps	Son	Plus de son	Praat
bfaDJ1g	bfaFC0	1	est-ce que tu peux me parler un peu de ta famille de ce que font tes parents et / frères et soeurs	0.001 - 4.96927			
bfaDJ1g	bfaDJ1	2	m (soupir) // donc ma mère est institutrice à l'école // communale de Montzen / en sixième primaire /	4.93469 - 14.10069			
bfaDJ1g	bfaDJ1	3	mon père lui il // travaille chez Coppland dans une usine / pour compresseur /	14.10069 - 20.56878			
bfaDJ1g	bfaDJ1	4	mais il va changer de	20.56878 - 22.06304			

transcription texte
et transcription PRAAT (alignée)

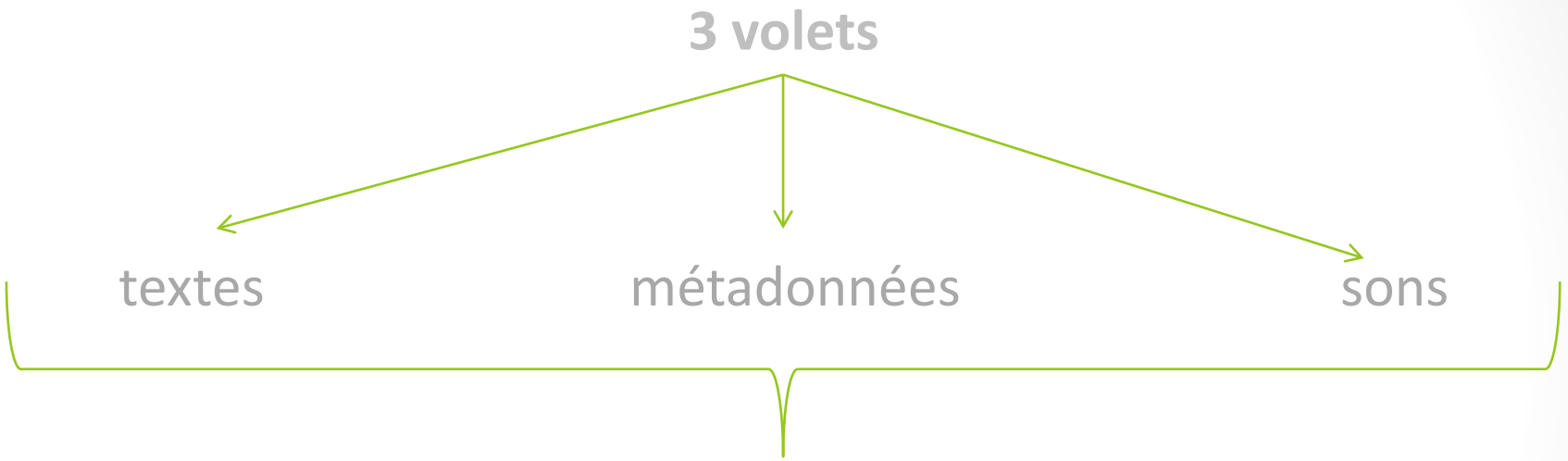
écoute du son correspondant à la ligne de transcription

écoute d'un extrait plus long (contexte)

noms de l'enregistrement
et du locuteur

n° de la ligne de transcription

base de données VALIBEL



interface de consultation **MOCA** (sous firefox)

http://moca.fltr.ucl.ac.be/moca/index.php

base de données VALIBEL

exploitation de la base de données et de ses corpus…

en **sociolinguistique** (insécurité linguistique)

chercheurs Valibel : Hambye, Romainville, Mariscal

L1 en Belgique je ne sais pas si euh // on on prend toujours comme modèle le français euh // de France (…) ben c' est quand-même le français vient de la France c' est quand-même le modèle // qu' il faut essayer de suivre

[VALIBEL, 1988, F, étudiante, 21 ans, Gilly (Hainaut)]

L1 oui parce-que on / ça oblige à parler // on on essaie de // de ne pas montrer qu' on vient de Belgique (rires de L0 et L1) (…) pour montrer qu' on / qu' on sait quand-même parler comme euh (…) comme les Français quoi

[VALIBEL, 1988, H, ingénieur technicien, 52 ans, Gilly (Hainaut)]

L1 ils (les Français) parlent mieux oui (…) bè i finissent mieux leurs phrases / i-y-a des mots qu' i disent mieux que nous |- hein <L0> mieux prononcés ? -| mieux prononcés que nous

[VALIBEL, 1988, F, commerçante, 60 ans, Roux (Hainaut)]

base de données VALIBEL

exploitation de la base de données et de ses corpus…

en **phonologie et prosodie** (projet PFC)

chercheurs Valibel : Bardiaux, Simon, Hambye

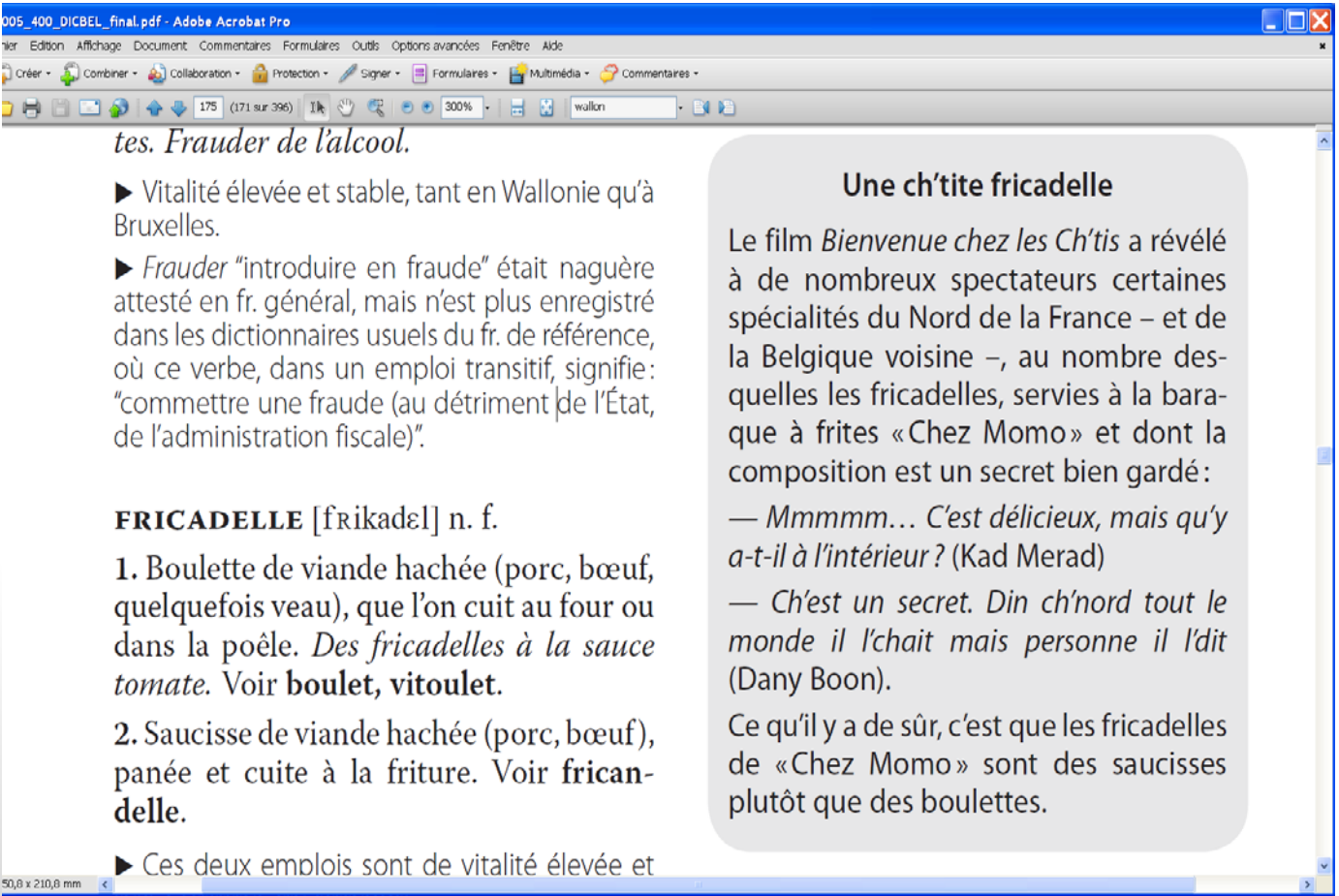
- 10 points d'enquête en Belgique francophone
- 12 locuteurs par point d'enquête
- 4 tâches : conversations libre et guidée, lectures mots et texte

http://www.projet-pfc.net/

base de données VALIBEL

exploitation de la base de données et de ses corpus…

en **en lexicologie différentielle** et **lexicographie**



tes. *Frauder de l'alcool.*

▶ Vitalité élevée et stable, tant en Wallonie qu'à Bruxelles.

▶ *Frauder* "introduire en fraude" était naguère attesté en fr. général, mais n'est plus enregistré dans les dictionnaires usuels du fr. de référence, où ce verbe, dans un emploi transitif, signifie : "commettre une fraude (au détriment de l'État, de l'administration fiscale)".

FRICADELLE [fʀikadɛl] n. f.

1. Boulette de viande hachée (porc, bœuf, quelquefois veau), que l'on cuit au four ou dans la poêle. *Des fricadelles à la sauce tomate.* Voir **boulet, vitoulet**.

2. Saucisse de viande hachée (porc, bœuf), panée et cuite à la friture. Voir **fricandelle**.

▶ Ces deux emplois sont de vitalité élevée et

Une ch'tite fricadelle

Le film *Bienvenue chez les Ch'tis* a révélé à de nombreux spectateurs certaines spécialités du Nord de la France – et de la Belgique voisine –, au nombre desquelles les fricadelles, servies à la baraque à frites « Chez Momo » et dont la composition est un secret bien gardé :

— *Mmmmm… C'est délicieux, mais qu'y a-t-il à l'intérieur ?* (Kad Merad)

— *Ch'est un secret. Din ch'nord tout le monde il l'chait mais personne il l'dit* (Dany Boon).

Ce qu'il y a de sûr, c'est que les fricadelles de « Chez Momo » sont des saucisses plutôt que des boulettes.

extrait du

Dictionnaire des Belgicismes

M. Francard, G. Geron, R. Wilmet, A. Wirth (2010)

base de données VALIBEL

exploitation de la base de données et de ses corpus…

en **analyse du discours**

chercheurs Valibel : Degand, Simon, Uygur, Crible, Grosman…

- repérage des phénomènes dans les transcriptions écrites
  > exportation des transcriptions : WorldSmith Tools, Intex, etc.
- mise en relation avec différents plans langagiers
  - segmentation de l'oral > **unités** de discours (cf. Degand & Simon, BDU)
  - **marqueurs** de discours > périphérie de la syntaxe, connecteurs causaux, reformulatifs, parenthétiques, marqueurs de subjectivité, grammaticalisation (cf. Degand)
  - gestion des **tours de parole** (cf. Simon)
  - marqueurs de **disfluence** (cf. Degand, ARC)
  - **structuration** et **typologie** textuelles (formel vs. informel, écrit vs. oral, etc.)

base de données VALIBEL

documenter la **diversité**

des **pratiques linguistiques**

sur le territoire **belge**

vs. ~~représenter le français parlé en Belgique~~

→ réservoir de données toujours ouvert

- masse **importante** mais **disparate** d'enregistrements
  > METADONNEES
- sous-ensembles de données plus **équilibrés**, plus **homogènes**, plus **spécifiques** > besoins de recherche précis → MOCA

au-delà du corpus : les enjeux de la variation

documenter la **diversité**

des **pratiques linguistiques**

sur le territoire **belge**

vs. ~~représenter le français parlé en Belgique~~

→ réservoir de données toujours ouvert

- masse importante mais disparate d'enregistrements
  > METADONNEES
- **sous-ensembles** de données plus **équilibrés**, plus **homogènes**, plus **spécifiques** > besoins de recherche précis → MOCA

au-delà du corpus : les enjeux de la variation

en fonction des **objectifs** et des **besoins** de recherche précis

- constitution de sous-corpus spécifiques
- à partir des métadonnées
  - genre, style : conversation, discours, débat, interviews, lecture…
  - profil sociolinguistique des locuteurs
  - contenu syntaxique, sémantique, lexical…

→ échantillonnage des données, contrôle des variables

au-delà du corpus : les enjeux de la variation

sous-corpus échantillonné > **représentativité**

- représentatif de quoi ?
- comment atteindre cette représentativité ?

une réflexion sur la notion même de variation et de variété est indispensable et constitue un préalable incontournable à toute étude sur corpus afin…

- d'identifier les **enjeux** et les éventuels **biais** méthodologiques
- d'assumer les **artefacts méthodologiques**
- d'interpréter les **résultats** dans ce cadre

→qu'est-ce que la variation?

→comment définir les variétés?

Discours VALIBEL Variation

Merci pour votre attention

La base de données textuelles orales **VALIBEL**

de la variation aux variétés

http://www.uclouvain.be/valibel.html

Alice Bardiaux (FNRS – UCL)