Annotation of plurilingual corpora

Experience from the CLAPOTY project

Pascal Vaillant
Université Paris 13
vaillant@univ-paris13.fr

International Research Days *Social Media*
Université Rennes 2
Rennes, 24/10/2015

CLAPOTY

This talk is about a project that aimed at collecting and analyzing a corpus of oral speech in situations of language contact.

I am presenting a collective work (detailed later).

The reference person for general information on the project is:

Isabelle Léglise <leglise@vjf.cnrs.fr>

Code-switching, code-mixing

- A well-known phenomena, especially in migrant communities

  ʕla:s liʔana *les moustiques* daba ʕrfti fajn huma mxbʕi:n taħt *le lit et là, donc pour pouvoir être sûr qu'il n'y a pas de moustiques... c'était la poubelle* xSSha tkun *vide* wtanqbD ana tanxarž hadši kullu *et* tandir lma *alors automatiquement* la kajn ši *moustique* hnaja mxbaʕ *elle cherche l'ombre, elle fout le camp* f *la journée*.

  (example from Bentahila & Davies, 1983)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Code-switching, code-mixing

- Also well-known as a side-effect of some specific socio-linguistic situations :
  - Diglossia (Ferguson 1959)
  - Cultural (incl. technical etc.) pressure from a dominant language (Thomason & Kaufmann 1988)
  - Language shift (*ibid.*)
  - Language attrition (*ibid.*)
  - Language death (*ibid.*)
  - Emerging Pidgins or Creoles (*ibid.*)

Code-switching, code-mixing

- Long regarded as a sociolinguistic topic, with no interest for descriptive linguistics
- For descriptive linguistics, noise or interference
- But: rising interest since the early 80s in the possibility to describe a “grammar” of code-switching (Sankoff & Poplack 1981; Joshi 1982; Bentahila & Davies 1983; Woolford 1983; Di Sciullo, Muysken & Singh 1986; Myers-Scotton & Azuma 1990; Myers-Scotton 1993)

Language contact: psycholinguistics

- What is happening in the head of a speaker switching between one language and another?
- Which constraints do the grammar of the two (or more) languages in contact impose on the actual productions of a plurilingual speaker?
- Understanding codeswitching phenomena is part of understanding utterance planning and production (Myers-Scotton 1993).

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Language contact: sociolinguistics

- Why do people mix different languages? Utilitarian issues may quickly be accounted for. All the other social and interactional functions of language mixing (why do *bilingual* people mix different languages) have to be studied.
- What do the participants of an interaction negociate when they switch languages?
- How does the environing society consider the different languages being used? How does it judge language mixing practices?

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Language contact: linguistics

- Language change is known to often be driven or induced by language contact

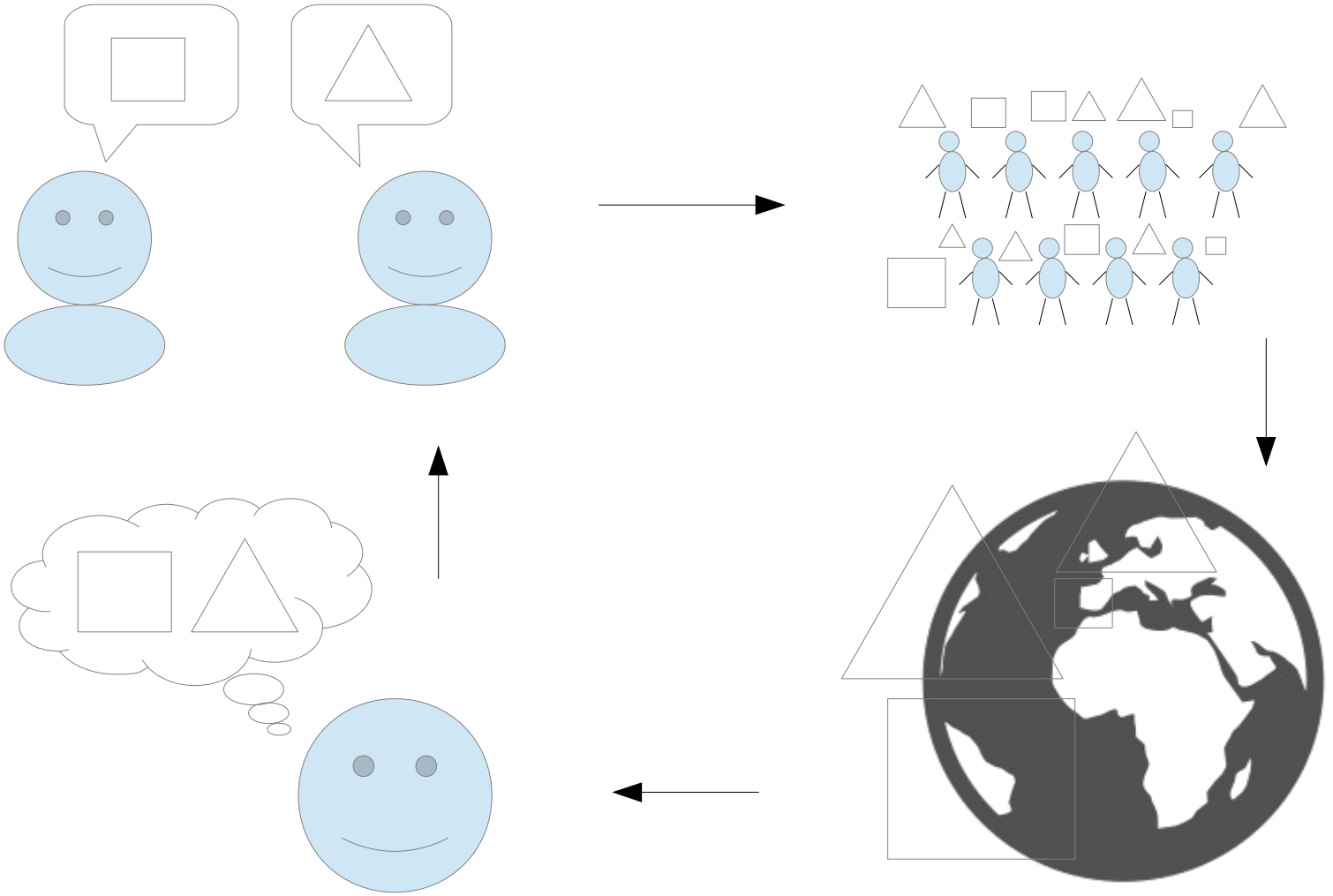
  (e.g. Old English with Norman French and Danish; Yiddish as a German dialect with Romance, Slavic and Hebrew elements; Creoles as heavily restructured European languages in the New World)

- Elementary steps of language evolution by contact must involve actual speakers of more than two languages interacting together

Language contact: linguistics

- Yet the methodology used to trace back language contact is the usual reconstruction method
- Intermediate stages seldom are documented: it is hard to hold both ends of the chain
- Understanding how language contact situations affect the languages being in contact is a key to understanding linguistic change (Thomason & Kaufmann 1988; Peyraube 2002; Heine & Kuteva 2005, 2007; Aikhenvald & Dixon 2006)

Language contact



Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

International Research Days *Social Media*
Université Rennes 2
Rennes, 24/10/2015

A corpus of live language contact

- 2009-2014: CLAPOTY Project (funded by the Agence Nationale pour la Recherche as ANR-09-JCJC-0121-01) : http://clapoty.vjf.cnrs.fr/
- Collect transcriptions of contemporary speech in language contact situations
- Develop computer tools to annotate, classify and search data
- See contact-induced language change in action

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

A corpus of live language contact

- 2009-2014: CLAPOTY: http://clapoty.vjf.cnrs.fr/
- Import knowledge from, and inform the fields of: sociolinguistics, typology, contact linguistics, formal linguistics, corpus linguistics
- Develop a multi-level and multifactorial methodology for description and analysis
- Develop computer standards to store and annotate plurilingual corpora and metadata
- Develop computer tools to mine plurilingual text (Léglise & Alby, 2013; Vaillant & Léglise, 2014)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Diversity

- 2009-2014: CLAPOTY: http://clapoty.vjf.cnrs.fr/
- A team of people with different scientific backgrounds

  Evangelia Adamou, Sophie Alby, Claudine Chamoreau, Anne Garcia-Fernandez, Gudrun Ledegen, **Isabelle Léglise**, Bettina Migge, Richard Nock, Claire Saillard, Duna Troiani, Pascal Vaillant
- Corpora displaying a great diversity of languages, *and* of language contact situations

Diversity of languages

- 40 languages, various typological features :
  - Native American: Kalin'a (French Guiana), Nahuatl, Purepecha (Mexico)
  - Creoles: French-based (Martinique, French Guiana, Réunion); English-based (Suriname); Portuguese-based (Guinea-Bissau)
  - West African: Wolof
  - West European: Romance languages (French, Portuguese, Spanish); Germanic languages (English, Dutch)
  - Balkan: Indo-European (Greek, Romani), Turkish
  - East-Asian: Austronesian (Aboriginal Taiwan languages: 'Amis and Truku); Chinese (Taiwan)

Diversity of situations

- Different language contact situations:
  - Stable plurilingualism (Purepecha & Spanish in Mexico; Truku & Chinese in Taiwan)
  - Creole continua (Martinique, Guiana, Réunion)
  - New emerging varieties (Suriname, Guiana)
- Different types of interaction:
  - Multiple participants: family (12), school (15), friends (24), media (15), work (51), interviews (27)
  - One speaker: political speech, narratives, tales

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Diversity of text plurilingualism

- Different degrees of internal heterogeneity:
  - near-monolingual texts, where the influence of language contact is felt through borrowings and typological changes in slow motion (Purepecha)
  - occasional code-switching (Guinea-Bissau Creole)
  - intensive code-switching, language mixing (Kalin'a)
  - fused lects (Turco-Romani)
  - Creole-Lexifier contacts within a continuum (Martinique, Réunion, Guiana)

(Auer 1999)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

The weight of descriptive frames

- Language contact phenomena have been described with different terms:
  - borrowing
  - code switching
  - intra-sentential code switching, code mixing
  - bilingual speech (*parler bilingue*)
  - fused lects, pidginization
  - interference, creolisms, *substratum* influence
  - calques, *pattern* or *matter* borrowing
  - etc.

The weight of descriptive frames

- Take a very common phenomenon, here described (on purpose) with no scientific words: an element of language A appears in an utterance of language B
  - the “element” may be a “word”, an idiom, a compound expression (possibly discontinuous), a complete utterance; it may be a system morpheme or a sequence of system morphemes;
  - even with no transfer of phonological matter, there may be prosodical features, semantic values, composition mechanisms... typical of B, used in A

The weight of descriptive frames

- Do you want to call this phenomenon *borrowing* or *code-switching*?
- The question seems pointless, but choosing either term has far-reaching implications on the conceptualization of what is actually happening

  e.g. implies different models:

  - different models of psycholinguistic processing
  - different models of (plurilingual) grammar
  - different models of language change mechanisms

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

The weight of descriptive frames

- What defines the limit between *borrowing* and *code-switching*? The size of the element? The "degree of integration" into the target language? What is it that defines a "degree of integration"? Frequency? Diachronical depth? Phonological integration?
- There has been long debates between specialists about what should be called "single-word code-switching" and what should be called "nonce borrowings" (Winford, 2003)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

A deliberately naive description

- We do not know with certainty who is right and who is wrong. We do not want to take sides.
- The use of some terms implies the use of some concepts; the use of some concepts implies adhering to a model
- There are some concepts that we do not wish to adopt without further inquiry, because they are subject to debate (e.g. *matrix language*)
- Structuring empirical data with *a priori* concepts the data is supposed to test would be illogical

Squaring the circle

- We want to: note, and annotate, all the possibly interesting language contact phenomena in a corpus, in order to be able to analyze them empirically
- We do not want to: use concepts that presuppose that these phenomena are already defined
- We need a new, multi-layer, annotation schema

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Annotating plurilingualism

- Why is it difficult?
- Let's take an example from CLAPOTY

(1.1)	**Yèr**	**mo**	**té**	**pasé**	**la**				
	hier	1SG	PST	passer	là				
	Hier je suis passé ici								
(1.2)	**i**	**té**	**gen**	***an***	madame	un peu	costaud	à côté	là
	3SG	PST	avoir	INDF	dame	un peu	costaud	à côté	là
	il y avait une dame un peu forte, à côté, là								
(1.3)	**i**	m'	a	donné	[. . .]	comme	***té***	***ni***	***problem***
	3SG	1SG	avoir	donner		comme	PST	avoir	problème
	elle m'a donné [. . .] *comme il y avait un problème*								

(Léglise/Nelson (2008) : *EDF* corpus – Cayenne)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Assigning a language to a word

- Several languages displayed
- Some of them share part of their lexical stock
- To the bilingual speakers, the question of whether they are picking the French *hier* or the Creole *yèr* does not arise
- To the linguists, there might be criteria to choose which language to assign a word to (e.g. phonological), but none is 100% certain

Finding the border of segments

- If a word, or sequence of words, belongs to the shared lexical stock of languages A and B

  (and hence may ambiguously been assigned to either of them)
- if there is a segment in language A *before* that word or sequence of words, and a segment in language B *after* it
- then where should we draw the border (on the syntagmatic axis) between A and B?

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Floating segments

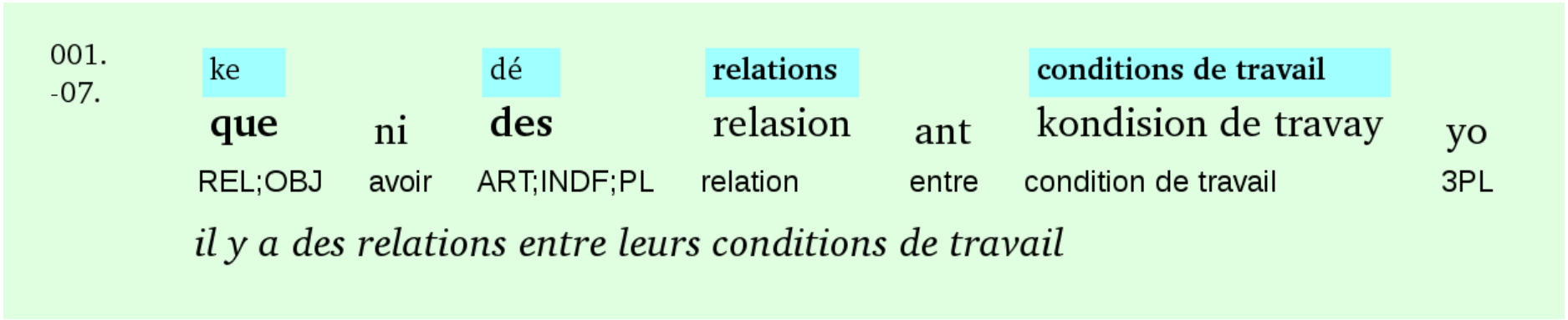
- Deciding to force the assignment of some words or segments to language A or B:
  - sometimes implies a near-random decision from the annotator, which yields uncertain data (minor sin)
  - always erases the actual complexity of the language contact situation (major sin)

⇒ Even choosing a transcription scheme is an arbitrary choice that imposes a grid on reality!

- Some segments simply “float” between languages (Ledegen, 2012)

Floating segments

- What we want to see is this:



Vaillant/Moustin (2007): *Voyé kriyé doktè ban mwen*

(Example displayed through the XSLT interface developed for CLAPOTY)

Implementation

- We want to be interoperable with other corpora
- we want to be state-of-the-art with regard to:
  - character encoding ⇒ Unicode
  - language encoding ⇒ BCP-47 (⊂ ISO-639)
  - document markup ⇒ XML
  - text annotation ⇒ TEI
- *but* we want our plurilingual segments
- and we want our language contact phenomena

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Beyond the TEI

- The *Text Encoding Initiative* has planned a lot of things, especially for corpora of oral transcripts (TEI-P5 Guidelines, chap. 8)

  … but it is somehow basic about how to describe linguistic heterogeneity:

  “Words or phrases which are not in the main language of the text should be tagged as such : ‘John eats a <foreign xml:lang="fr">croissant</foreign> every morning’.” (TEI-P5 Guidelines, p. 65)

- That’s ½ page in 1600 pages.

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

What's in a language tag?

- BCP-47:

  language tag (ISO-639)
  + optional variant tag (IANA Language subtags registry)
  + optional script tag (ISO-15924)
  + optional tag for geographic variant (ISO-3166 country codes, or UN M49 zone codes)

- Examples:
  - fra *vs.* fra-GF
  - spa *vs.* spa-419
  - djk-aluku *vs.* djk-ndyuka

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

What's in a language tag?

- ISO-639-3 also has tags for macrolanguages, if needed (e.g. ara, que, zho)
- There is an implicit hierarchy of specificity in language identification (zho > cmn > cmn-TW (to be used with caution)
- ISO-639-3 also has three tags with "special values":
  - 'und' : undefined
  - 'zxx' : no linguistic content
  - 'mul' : multiple languages

Beyond the TEI

- We want to be TEI-conformant as much as possible...

  and to create our own extensions when needed

- So we created a *XML Schema* of *Documents* (XSD) adapted to our needs: **Corpus-Contacts**

  (XSDs are like DTDs except that they also allow to specify integrity constraints)

- Essentially based on TEI-P5 Guidelines chap. 8 (*Transcriptions of speech*)

  … plus some new element types

The Corpus-Contacts XSD

- General structure :
  - The root element is a <corpus>
  - a <corpus> contains one <corpus_header>, then an indefinite number (1..n) of texts (elements <text>)
  - a <text> contains one <text_header>, then an indefinite number of events (elements <event>)
  - an <event> may be either a paraverbal element (<incident>, <kinesic>, <vocal>), or a speech turn
  - a <speech_turn> consists in four tiers: transcription, interlinear morphemic gloss, list of POS-tags, free translation

The Corpus-Contacts XSD

- Inside a transcription: TEI-P5 elements:
  - plain UTF-8 text (#PCDATA)
  - alignment tabs (to align with the IMG & POS-tags)
  - paraverbal events (incident, kinesic, vocal)
  - linguistic indications (shifts in pitch, tempo, loudness, rhythm, tension, voice quality)
  - pauses
  - overlaps
  - incomplete forms

The Corpus-Contacts XSD

- Specific **Corpus-Contacts** elements:
  - internal plurilingualism:
    - assignment to multiple languages
    - alternate transcriptions in multiple languages
  - remarkable phenomena

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Multiple language assignment

- The basic idea: when a segment is multilingual:
- continue using the basic **xml:lang** attribute (backward compatibility)
- give it value “mul” (ISO-639-3 special tag: *Multiple languages*)
- add a new element <langues> to give the list of alternate languages it is “floating among”:

```
<langues>
  <langue xml:lang="fra">
  <langue xml:lang="acf">
</langues>
```

Multiple language assignment (P)

- What does “multilingual segment” mean?

  (P) **Paradigmatic** interpretation: when the segment, with similar phonetic forms in A and B, does not give enough hints as to what language it should be assigned to (A or B),

  then tagging it “mul” means: this could be A and this could also be B (I, linguist, don’t know);

  or: this could be some linguistic item floating between A and B in bilingual speech.

- A and B are specified in the <langues> element

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Multiple language assignment (P)

- What does “multilingual segment” mean?

  (P) **Paradigmatic** interpretation:

		pour	l’instant					
(2)	***Piské***	***pou***	***lenstan***	***sé***	***journalis***	***ki***	***ni***	***la***
	puisque	pour	l’instant	être.COP	journaliste	REL;SBJ	avoir	là

Puisque pour l’instant ce (ne) sont (que) des journalistes qui sont là

(Vaillant/Lengrai (2007) : *Lignes de Vie*)

Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Multiple language assignment (P)

```
<transcription lang="acf">
piskè <tab/>
 
  <langues><langue lang="acf"/><langue lang="fra"/></langues>
   <trans_alt lang="acf">pou</trans_alt> <trans_alt
lang="fra">pour</trans_alt> <tab/>
   <trans_alt lang="acf">lenstan</trans_alt>
<trans_alt lang="fra">l'instant</trans_alt>
  <tab/>
 sé <tab/> jounalis <tab/> ki <tab/> ni <tab/> la
```

Multiple language assignment (S)

- What does “multilingual segment” mean?

  (S) **Syntagmatic** interpretation: when a segment comprising multiple morphemes (in A and in B, and/or in subsegments floating between A and B) does not allow to clearly identify a main language which rules the syntagm construction,

  then tagging it “mul” means: this segment is internally multilingual (no “matrix language”)
- A and B are specified in the <langues> element

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Multiple language assignment (S)

- What does “multilingual segment” mean?

  (S) **Syntagmatic** interpretation:

(3.1)	Ah	oui	mais	même	si	**ou**	**ka**	**vin**,
	INTJ	oui	mais	même	si	2SG	IPFV	venir
	Ah oui, mais même si vous venez,							
(3.2)	tant	**ou**	**pa**	***ni***	**tout**		**papié**	**a**
	tant	2SG	NEG	avoir	tout.QUANT		papier	DEF
	tant que vous n’avez pas tous les papiers ...							

(Léglise/Nelson (2008) : *EDF* corpus – Cayenne)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Multiple language assignment (S)

```
<transcription lang="mul">
 <langues><langue lang="gcr"/><langue lang="fra"/><langue
lang="acf"/></langues>
 Ah <tab/> oui <tab> mais <tab/> même
<tab/> si <tab/>
 ou <tab/> ka <tab/> vin <tab/>
 tant <tab/>
 ou <tab/> pa <tab/>
 ni <tab/>
 tout <tab/> papié <tab/> a
```

Multiple language assignment (PS)

- Of course both interpretations are possible at the same time (this is the case with a segment fragmented in several subsegments which themselves are “floating” units):

Vini	**non**	bande	de	putes
venir	d’accord	bande	de.GEN	pute

Venez ici, bande de putes

Multiple language assignment (PS)

```
<transcription lang="mul">
<langues><langue lang="gcr"/><langue lang="fra"/></langues>
 vini
 <tab/>
 <langues><langue lang="gcr"/><langue
lang="fra"/></langues>
  non
 <tab/>
 bande <tab/> de <tab/> putes
```

Remarkable phenomena

- Another purposefully naive description: we refuse to use predefined categories of language contact phenomena
- Remarkable phenomenon: “something is worth noting here”
- Just a generic frame to annotate everything worth analyzing

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Remarkable phenomena

- We use the generic element <passage_remarquable> (remarkable passage) to signal the occurrence of a remarkable phenomenon somewhere in a text in the corpus
- Every “remarkable passage” has an XML ID tag
- In the database, remarkable passages (tokens) are linked to remarkable phenomena (types)
- An indefinite number (1..n) of remarkable passages may be linked to a single remarkable phenomenon

Remarkable phenomena

- In the database, there is a hierarchy of remarkable phenomena
- The predefined description levels are not linked to a theoretical model of language contact, but are data-oriented: they specify (I) which layer of language processing is involved; (II) which type of syntagm is affected
- The last description level is meant to be created and maintained in a bottom-up process by linguists users of the database

Meta-categories of R.Ph.

- First level of the hierarchy : three main meta-categories : PREMS, PRINT and PREDISC
  - **PREMS** :
    Phénomènes REmarquables Morpho Syntaxiques
  - **PRINT** :
    Phénomènes Remarquables INTeractionnels
  - **PREDISC** :
    Phénomènes REmarquables DISCursifs

PREMS

- Morphosyntactically remarkable phenomena
- Tactical subtypes: defined by the position of the remarkable phenomenon ([]) in the chain of alternating language segments (<A><B>)
- Symbolic notation for the four tactical subtypes:

[<>]	the presence of a segment of B in A is remarkable
[<><>]	the sequence of two segments in languages A and B is remarkable
<[><]>	the switch between A and B is remarkable
<[]>	something inside language A is remarkable

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

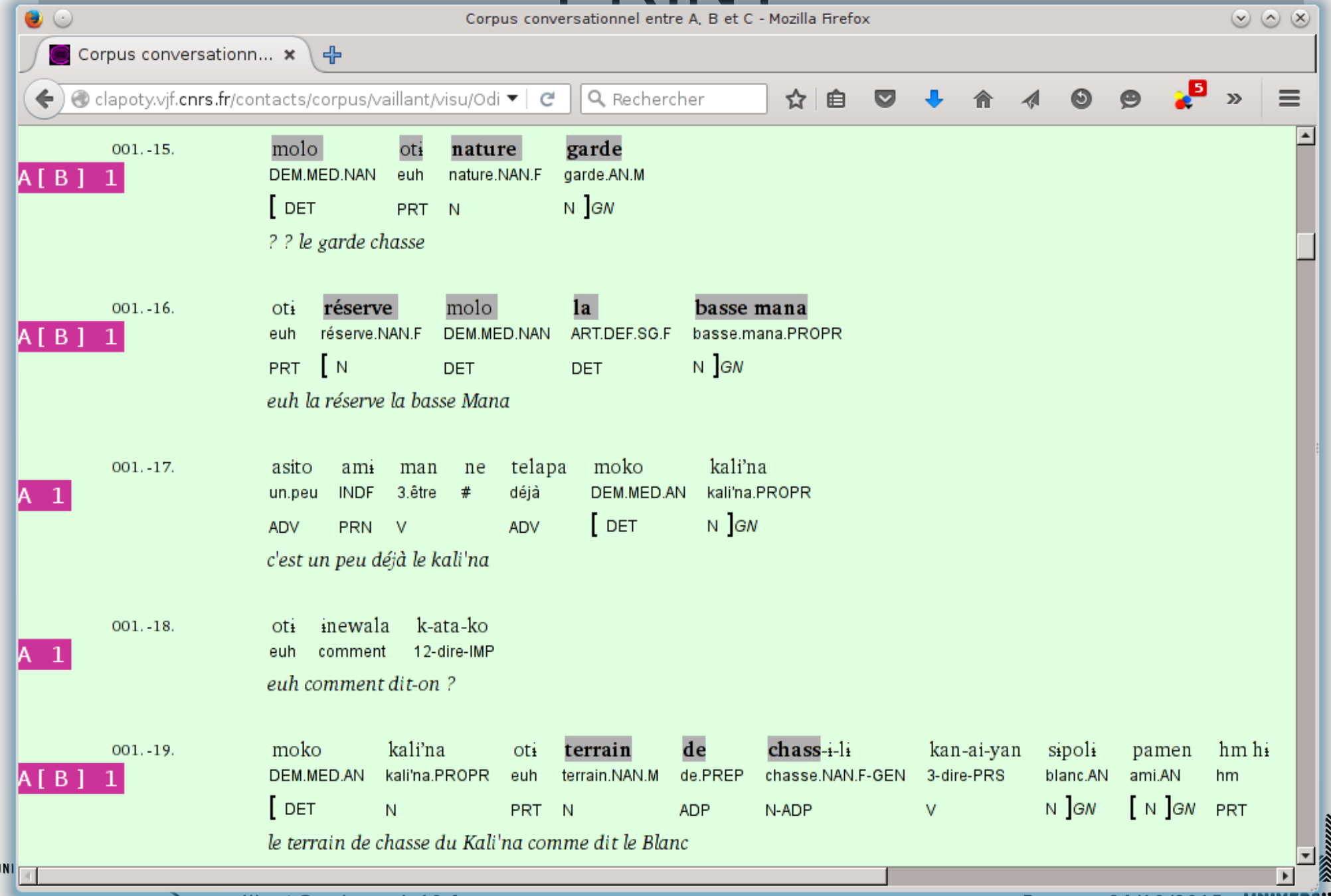
PREMS

- Morphosyntactically remarkable phenomena
- Subcategories under the major tactical subtypes: defined by the type of syntagm affected:
  - PREMS-GV: in the Verb Phrase
  - PREMS-GN: in the Noun Phrase
    - PREMS-GN-Det : concerning determination in the NP
    - PREMS-GN-Poss : concerning the expression of possession in the NP
  - etc.

PRINT

- Concerns the analysis of the alternation of languages w.r.t. speakers during the interaction (Auer, 1995)
- A preliminary automatic annotation “à la Auer” (Language A [Language B] – Speaker 1) is automatically computed by the XSLT processor

PRINT


Corpus conversationnel entre A, B et C - Mozilla Firefox
Corpus conversationn...
clapoty.vjf.cnrs.fr/contacts/corpus/vaillant/visu/Odi
Rechercher
001.-15.
A [B] 1
molo oti nature garde
DEM.MED.NAN euh nature.NAN.F garde.AN.M
[DET PRT N N]GN
? ? le garde chasse
001.-16.
A [B] 1
oti réserve molo la basse mana
euh réserve.NAN.F DEM.MED.NAN ART.DEF.SG.F basse.mana.PROPR
PRT [N DET DET N]GN
euh la réserve la basse Mana
001.-17.
A 1
asito ami man ne telapa moko kali'na
un.peu INDF 3.être # déjà DEM.MED.AN kali'na.PROPR
ADV PRN V ADV [DET N]GN
c'est un peu déjà le kali'na
001.-18.
A 1
oti inewala k-ata-ko
euh comment 12-dire-IMP
euh comment dit-on ?
001.-19.
A [B] 1
moko kali'na oti terrain de chass-i-li kan-ai-yan sipoli pamen hm hi
DEM.MED.AN kali'na.PROPR euh terrain.NAN.M de.PREP chasse.NAN.F-GEN 3-dire-PRS blanc.AN ami.AN hm
[DET N PRT N ADP N-ADP V N]GN [N]GN PRT
le terrain de chasse du Kali'na comme dit le Blanc

PREDISC

- Concerns the impact of plurilingualism on discourse cohesion and articulation
- e.g. discourse connectors imported from another language in situations of cultural pressure

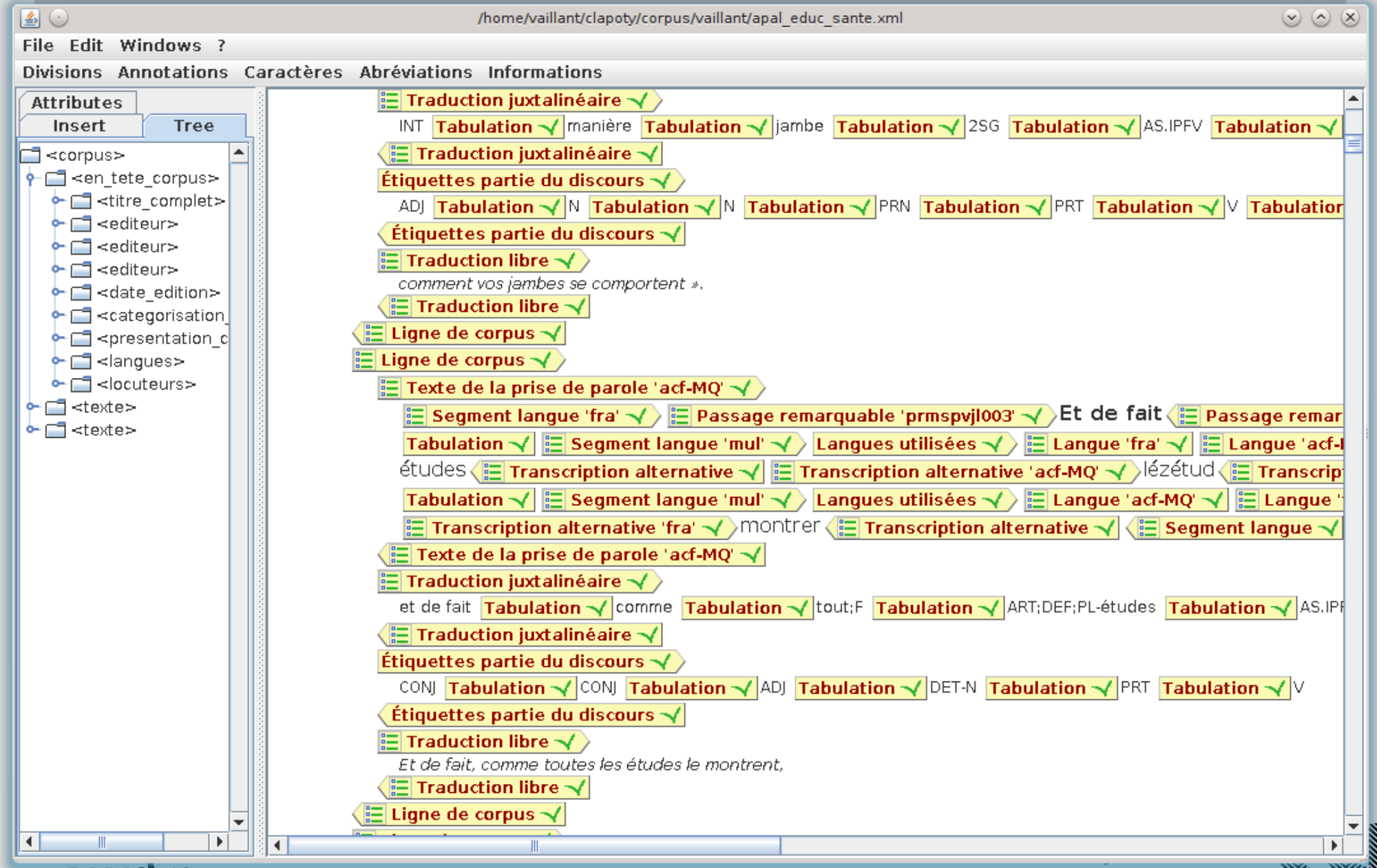
UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set

- The XSD Document Schema Corpus-Contacts
- A specific config file for the open-source java-based JAXE XML editor

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set



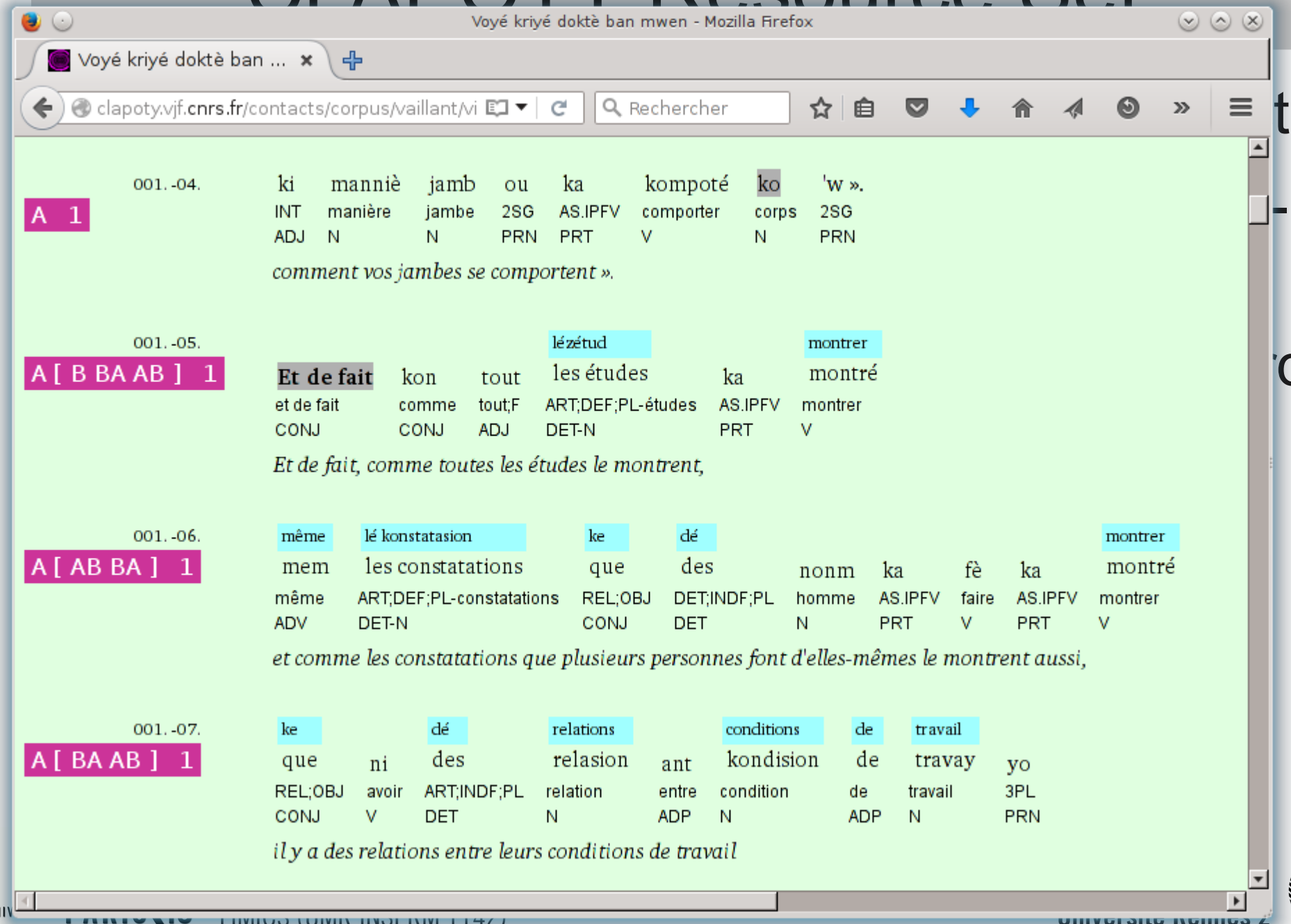
UNIVERSITÉ PARIS 13
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set

- The XSD Document Schema Corpus-Contacts
- A specific config file for the open-source java-based JAXE XML editor
- A XSLT transform sheet allowing any standard XSLT-1.0 conformant browser to display the corpora as a sequence of aligned utterances

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set



001. -04.

A 1

ki	manniè	jamb	ou	ka	kompoté	ko	'w ».
INT	manière	jambe	2SG	AS.IPFV	comporter	corps	2SG
ADJ	N	N	PRN	PRT	V	N	PRN

comment vos jambes se comportent ».

001. -05.

A [B BA AB] 1

			lézétud		montrer
Et de fait	kon	tout	les études	ka	montré
et de fait	comme	tout;F	ART;DEF;PL-études	AS.IPFV	montrer
CONJ	CONJ	ADJ	DET-N	PRT	V

Et de fait, comme toutes les études le montrent,

001. -06.

A [AB BA] 1

même	lé konstatasion		ke	dé					montrer
mem	les constatations		que	des	nonm	ka	fè	ka	montré
même	ART;DEF;PL-constatations		REL;OBJ	DET;INDF;PL	homme	AS.IPFV	faire	AS.IPFV	montrer
ADV	DET-N		CONJ	DET	N	PRT	V	PRT	V

et comme les constatations que plusieurs personnes font d'elles-mêmes le montrent aussi,

001. -07.

A [BA AB] 1

ke		dé	relations		conditions	de	travail	
que	ni	des	relasion	ant	kondision	de	travay	yo
REL;OBJ	avoir	ART;INDF;PL	relation	entre	condition	de	travail	3PL
CONJ	V	DET	N	ADP	N	ADP	N	PRN

il y a des relations entre leurs conditions de travail

vaillant@univ-paris13.fr

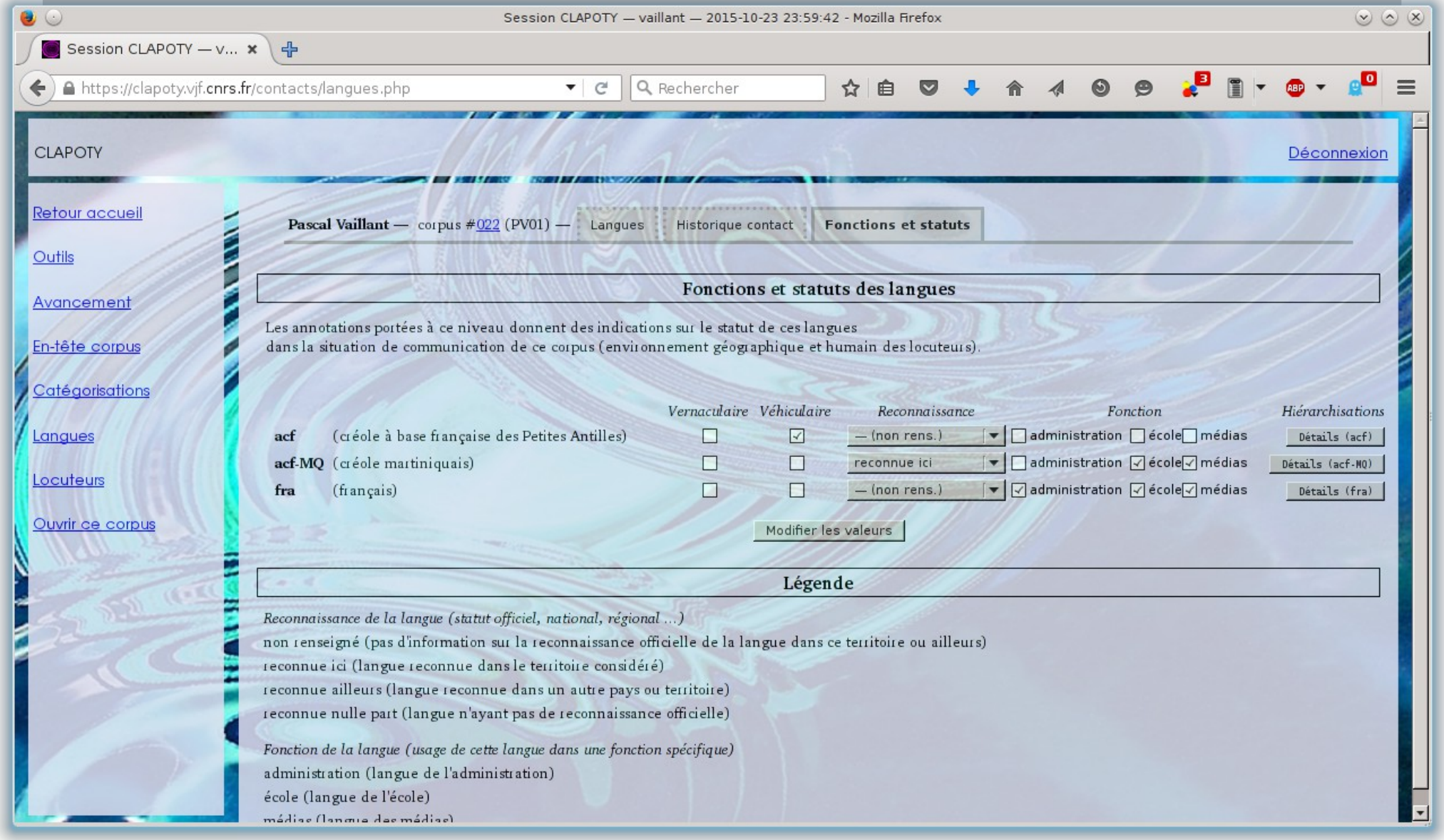
Rennes, 24/10/2015

CLAPOTY Resource Set

- The XSD Document Schema Corpus-Contacts
- A specific config file for the open-source java-based JAXE XML editor
- A XSLT transform sheet allowing any standard XSLT-1.0 conformant browser to display the corpora as a sequence of aligned utterances
- A relational (SQL) database to store sociolinguistic information on corpora, speakers, languages

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set



UNIVERSITÉ PARIS 13

Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

International Research Days *Social Media*
Université Rennes 2
Rennes, 24/10/2015

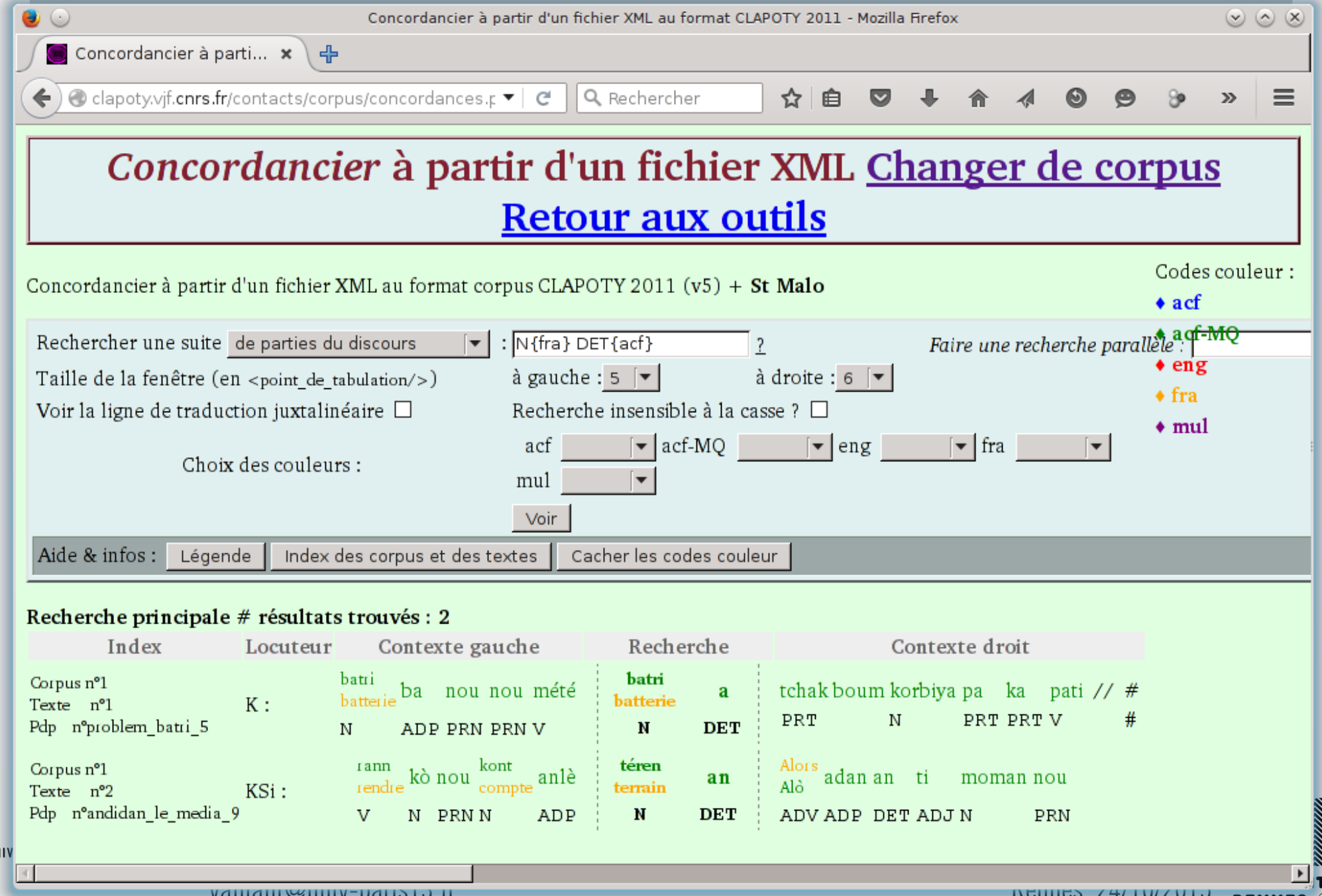
UNIVERSITÉ RENNES 2

CLAPOTY Resource Set

- The XSD Document Schema Corpus-Contacts
- A specific config file for the open-source java-based JAXE XML editor
- A XSLT transform sheet allowing any standard XSLT-1.0 conformant browser to display the corpora as a sequence of aligned utterances
- A relational (SQL) database to store sociolinguistic information on corpora, speakers, languages
- A concordancer to search for patterns

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

CLAPOTY Resource Set



Concluding Remarks

- Relevance to Network-Mediated Communication?

Plurilingualism in written form

- Language mixing is not limited to oral speech
- Oldest written testimony in transcripts from Martin Luther:

  si enim hoc verum esset, so schiss ich dem pabst auf die kron.

  (example from Stolt 1964, quoted in Auer & Muhamedova 2005)
- “Oralized” writing

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Plurilingualism in written form

- With instant messaging systems (IM, SMS...) and more generally CMC, there is a wealth of new types of communication which:
  - are written;
  - are no oral transcript or oralized writing;
  - yet differ from what used to be considered written language in many parameters.
- Some of these new forms of communication exhibit internal language mixing.

Language mixing on social networks

UNIVERSITÉ PARIS 13

Pascal V
LIMICS (
vaillant@

Research Days *Social Media*
Université Rennes 2
Rennes, 24/10/2015

UNIVERSITÉ RENNES 2

Language mixing in UGC (forums)

[redacted]
ma fierté c mes deux enfants

Posté le 01-09-2008 à 23:39:53

saha ftourkom pour celle font le ramdam; bonsoir tout le monde je viens juste vous dire que j'ai perdu 2 petit kg. bon courage toute et bien venue aux autres

J'aime 0

[redacted]
Rien n'est Impossible
Profil : Fidèle

Posté le 02-09-2008 à 00:22:28

sa7a chribtek toi aussi y3aychek

mais waw 2 kilo c super j'espère que tu contiura endant ramadan malgré que c dur
on vas y arriver pourtant

[redacted] 23 ans
debut :le 3 aout 2008 avec 90Kg AIE
---- 25/08/08 : 87KG ---- 29/08/08 : 84KG ---- 20/09/08 : 82KG ----------

11/10/08 : 78KG ------ 12/11/75 : 75KG -------- 20/12/08 : 72KG ---- objectif : 70KG pour 05/01/09 pfff:

Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Language mixing in SMS

- “Ok pour le pot ! Suis 3 les 2 3 et 4as”

  (A friend of mine, p.c.)
- Cf. Simone Ueberwasser’s talk about sms4science.ch

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

Open Questions

- In Network-Mediated Communication:
- The issue of plurality of languages exists
- It interacts with other issues:
  - plurality of writing systems and encodings
  - plurality of writing standards
  - plurality of *genres* and genre-specific varieties
  - variable levels of conformance to writing standards

    (at the speech community level, age/occupation group level, user community level, individual level)

UNIVERSITÉ PARIS 13
Pascal Vaillant
LIMICS (UMR INSERM 1142)
vaillant@univ-paris13.fr

References

- A. Bentahila & E.E. Davies, “The syntax of Arabic-French code-switching”. *Lingua* 59 (1983), p. 301-330.
- C. Ferguson, “Diglossia”. *Word* 15 (1959), p. 325-340.
- S. Thomason & T. Kaufmann, *Language Contact, Creolization, and Genetic Linguistics*. University of California Press (1988).
- D. Sankoff & S. Poplack, “A formal grammar for code-switching”. *Papers in Linguistics* 14 (1981), p. 3-46.
- A. Joshi, “Processing of sentences with intra-sentential code-switching”. COLING 1982 (Prague).
- E. Woolford, “Code-switching and syntactic theory”. *Linguistic Inquiry* 14 (3) (1983), p. 520-536.

References

- A.-M. Di Sciullo, P. Muysken & R. Singh, “Government and code-mixing”. *Journal of Linguistics* 22 (1986), p. 1-24.
- C. Myers-Scotton & S. Azuma, “A frame-based process model of codeswitching”. 26th annual regional meeting of the *Chicago Linguistic Society* (1990).
- C. Myers-Scotton, *Duelling Languages : Grammatical Structure in Codeswitching*. Oxford University Press (1993).
- A. Peyraube, “L’évolution des structures grammaticales”. *Langages* 146 (2002), p. 46-58.
- B. Heine & T. Kuteva, *Language Contact and Grammatical Change*. Cambridge University Press (2005).
- B. Heine & T. Kuteva, *The Genesis of Grammar*. Oxford University Press (2007).

References

- A. Aikhenvald & R. M. Dixon (eds.), *Grammars in Contact: A Cross-Linguistic Typology*. Oxford University Press (2006).
- I. Léglise & S. Alby, “Les corpus plurilingues, entre linguistique de corpus et linguistique de contact”. *Faits de langues* 41 (2013), p. 95-122.
- P. Vaillant & I. Léglise, “À la croisée des langues : Annotation et fouille de corpus plurilingues”. *RNTI SHS* 2 (2014), p. 81-100.
- P. Auer, “From code-switching via language mixing to fused lects: Toward a dynamic typology of bilingual speech”. *International Journal of Bilingualism* 3 (4) (1999), p. 309-332.
- D. Winford, *An Introduction to Contact Linguistics*. Blackwell (2003).

References

- G. Ledegen, “Prédicats ‘flottants’ entre le créole acrolectal et le français à la Réunion : Exploration d’une zone ambigüe”. In C. Chamoreau & L. Goury (eds.): *Changements linguistiques et langues en contact : Approches plurielles du domaine prédicatif*. CNRS Éditions (2012), p. 251-270.
- P. Auer, “The pragmatics of code-switching: a sequential approach”. In L. Milroy & P. Muysken (eds.): *One Speaker, Two Languages: Cross-Disciplinary Perspectives on Code-Switching*. Cambridge University Press (1995), p. 115-135.
- B. Stolt, *Die Sprachmischung in Luthers Tischreden*. Stockholm: Almqvist & Wiksell (1964).
- P. Auer & R. Muhamedova, “‘Embedded language’ and ‘matrix language’ in insertional language mixing: Some problematic cases”. *Rivista di Linguistica* 17.1 (2005), p. 35-54.