Annotation of plurilingual corpora

Experience from the CLAPOTY project

Pascal Vaillant
Université Paris 13
vaillant@univ-paris13.fr
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>
A well-known phenomena, especially in migrant communities

ʕlaːs liʔana les moustiques daba ʕrfti fajn huma mxbʕiːn 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)
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).
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 negotiate when they switch languages?

• How does the environing society consider the different languages being used? How does it judge language mixing practices?
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
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
A corpus of live language contact

- 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)
Diversity


- 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
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)
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
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)
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
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 problème
3SG 1SG avoir donner comme PST avoir problème
elle m’a donné [...] comme il y avait un problème

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

001. -07. 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

\[ il\ y\ a\ des\ relations\ entre\ leurs\ conditions\ de\ travail \]

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 $\Rightarrow$ Unicode
  - language encoding $\Rightarrow$ BCP-47 ($\subset$ ISO-639)
  - document markup $\Rightarrow$ XML
  - text annotation $\Rightarrow$ TEI
- *but* we want our plurilingual segments
- and we want our language contact phenomena
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.
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
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
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”:

    ```xml
    <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
What does “multilingual segment” mean?

\textbf{(P) Paradigmatic interpretation:}

\begin{verbatim}
pour l’instant

(2) Piské pou lenstan sé journalis ki ni la
puisque pour l’instant être.COP journalistes REL;SBJ avoir là
Puisque pour l’instant ce (ne) sont (que) des journalistes qui sont là
\end{verbatim}

(Vaillant/Lengrai (2007) : \textit{Lignes de Vie})
<transcription lang="acf">
piskè <tab/>
  <segment lang="mul">
    <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>
  </segment> <tab/>
  sé <tab/> jounalis <tab/> ki <tab/> ni <tab/> la
</transcription>
<traduction_juxtalineaire>
  puisque <tab/> pour <tab/> l'instant <tab/> être.COP <tab/> journaliste <tab/> REL;SBJ <tab/> avoir <tab/> là
</traduction_juxtalineaire>
<traduction_libre>
  puisque pour l'instant ce (ne) sont (que) des journalistes qui sont là,
</traduction_libre>
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
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)
Ah oui, mais même si vous venez, tant que vous n'avez pas tous les papiers …
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>
<segment lang="gcr">vini</segment>
<tab/>
<segment lang="mul">
<langues><langue lang="gcr"/>
<langue lang="fra"/>
</langues>
non</segment>
<tab/>
<segment lang="fra">bande <tab/> de <tab/> putes</segment>
</transcription>
<traduction_juxtalineaire>
venir <tab/> d'accord <tab/> bande <tab/> de.GEN <tab/>
pute.PL
</traduction_juxtalineaire>
<traduction_libre>venez ici, bande de putes</traduction_libre>
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
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
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.
• 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
Concerns the analysis of the alternation of languages w.r.t. speakers during the interaction (Auer, 1995).

```
001.15.  molo  oti  nature  garde
         DEM.MED.NAN  euh  nature.NAN.F  garde.AN.M
         [ Det  PRT  N  N ]

? ? le garde chasse

001.16.  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 ]

euh la réserve la basse Mana

001.17.  asto  ami  man  ne  telap a moko  kalì'na
         un.peu  INDF  3.ère  #  déjà  DEM.MED.AN  kalì'na.PROPR
         ADV  PRN  V  ADV  [ Det  N ]

c'est un peu déjà le kalì'na

001.18.  oti  inewala  k-ata-ko
         euh  comment  12-dire-IMP
         [ N  V ]
euh comment dit-on ?

001.19.  moko  kalì'na  oti  terrain  de  chass-i-li  kan-ai-yam  sipoli  pamen  hm li
         DEM.MED.AN  kalì'na.PROPR  euh  terrain.NAN.M  de.FREP  chasse.NAN.F-GEN  3-dire-PRS  blanc.AN  amil.AN  hm
         [ Det  N  PRT  N  ADP  N-ADP  V  N ]

le terrain de chasse du Kalì'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
CLAPOTY Resource Set

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

001.-04.

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

comment vos jambes se comportent ».

001.-05.

Et de fait kon tout les études ka montrer
et de fait comme toutF 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.

mème le konstatasion ke dé
mème les constatations que des
ADV DET-N ART;DEF;PL-constatations REL;OBJ CONJ DET;INDF;PL DET N

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

001.-07.

ke dé relations conditions de travail
que ni des relasion ant kondision de travail
REL;OBJ V DET CONJ N ART;INDF;PL DET ADP N

il y a des relations entre leurs conditions de travail
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
CLAPOTY Resource Set

Les annotations portées à ce niveau donnent des indications sur le statut de ces langues dans la situation de communication de ce corpus (environnement géographique et humain des locuteurs).

**Fonctions et statuts des langues**

- **Vernaculaire**
- **Véhiculaire**
- **Reconnaissance**
- **Fonction**
- **Hiérarchisations**

<table>
<thead>
<tr>
<th>Langue</th>
<th>Description</th>
<th>Vernaculaire</th>
<th>Véhiculaire</th>
<th>Reconnaissance</th>
<th>Fonction</th>
<th>Hiérarchisations</th>
</tr>
</thead>
<tbody>
<tr>
<td>aef</td>
<td>créole à base française des Petites Antilles</td>
<td>☑️</td>
<td>☑️</td>
<td>(non reconnue)</td>
<td>☑️ administration ☑️ école ☑️ médias</td>
<td>Détails (aef)</td>
</tr>
<tr>
<td>aef-MQ</td>
<td>créole martiniquais</td>
<td>☑️</td>
<td></td>
<td>reconue ici</td>
<td>☑️ administration ☑️ école ☑️ médias</td>
<td>Détails (aef-MQ)</td>
</tr>
<tr>
<td>fra</td>
<td>français</td>
<td>☑️</td>
<td></td>
<td>(non reconnue)</td>
<td>☑️ administration ☑️ école ☑️ médias</td>
<td>Détails (fre)</td>
</tr>
</tbody>
</table>

**Légende**

- Reconnaissance de la langue (status officiel, national, régional...)
- non reconnue (pas d'information sur la reconnaissance officielle de la langue dans ce territoire ou ailleurs)
- reconnue ici (langue reconnue dans le territoire considéré)
- reconnue ailleurs (langue reconnue dans un autre pays ou territoire)
- reconnue nulle part (langue n'ayant pas de reconnaissance officielle)

- Fonction de la langue (usage de cette langue dans une fonction spécifique)
  - administration (langue de l'administration)
  - école (langue de l'école)
  - médias (langue des médias)
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
CLAPOTY Resource Set

Concordancier à partir d'un fichier XML

Changer de corpus

Retour aux outils
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
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
Language mixing in UGC (forums)

**User 1:** ma fierté c mes deux enfants

Posté le 01-09-2008 à 23:39:53

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

**User 2:** Rien n'est Impossible

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 😊

**User 3:** 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 pfft:
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
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)
References

References

References


References


