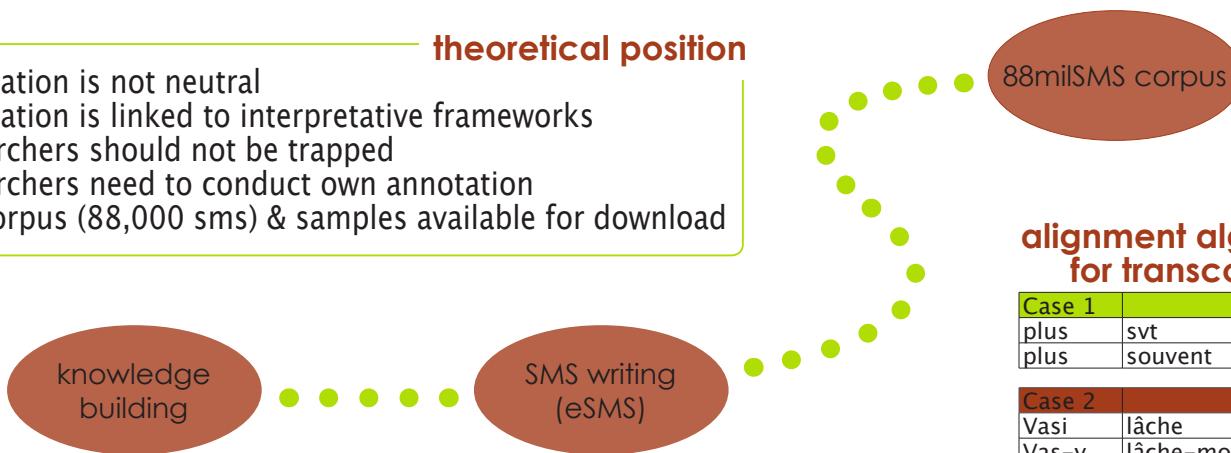


# Non-standard texts: from theoretical positions to Natural Language Processing normalisation

- theoretical position**
- > annotation is not neutral
  - > annotation is linked to interpretative frameworks
  - > researchers should not be trapped
  - > researchers need to conduct own annotation
  - > full corpus (88,000 sms) & samples available for download



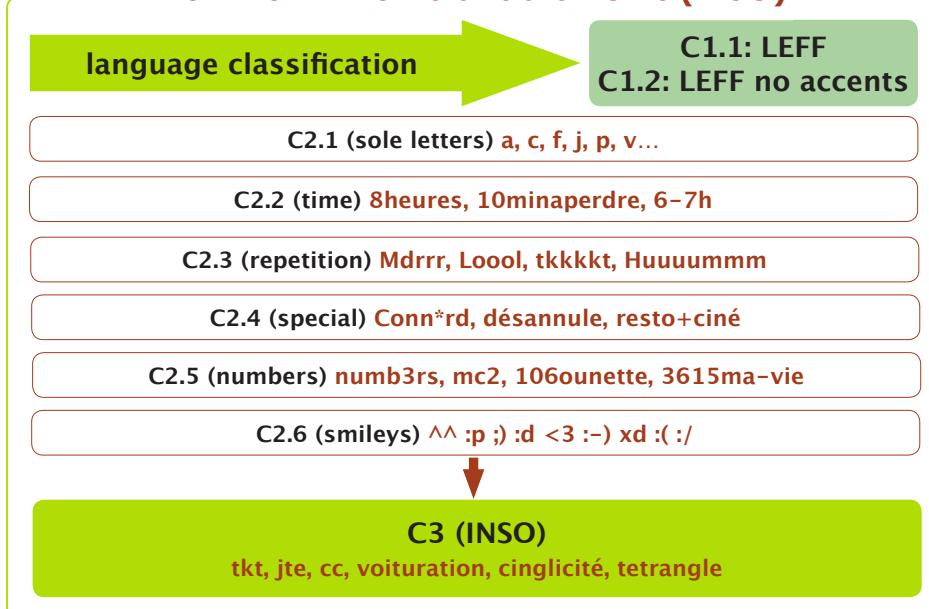
## alignment algorithms for transcoding

Case 1		
plus	svt	possible
plus	souvent	possible

Case 2		
Vasi	lâche	moi
Vas-y	lâche-moi	

Case 3		
T'as		eu
Tu	as	eu

## 'unknown' non-standard items (INSO)



## automatic normalisation techniques

- > new typology of detected 'mistakes'
- > normalisation based on most frequent errors
  - > confrontation with:
    - traditional automatic translation,
    - speech recognition,
    - spelling/grammatical checker principles
- > comparison between different types of instant media (SMS, forums, tweets)

# substitution

	entire (S.P.E): o (eau), 7 (cet)	1.
phonetic	partial: ossi (aussi), allé (aller), bizes (bises)	2.
	with variation (S.P.V): k (que), kikou (coucou)	3.
graphical	elision, typography, capitals/lower case: m en, est ce que	4.
	icons, symbols, rebuses: à + (à plus), de grandes @ (oreilles)	5.
	with variation: bisoux (bisous), mwa (moi)	6.

# reduction

	morpho-lexical shortenings: initialisms - alphabetisms & acronyms: ASV, mdr, tvb, tlm, lol truncations - apocope; aphaeresis: ordi (ordinateur), 'lut, Net (salut, Internet)	7. a) & 7. b)
phonetic	entire: c (c'est/ces), d(des/dé/dès), v (vais)	7. c), d), e)
	variation: ui (oui)	8.
graphical	suppression of mute word-endings; word-beginnings: vou (vous), peu (peut), ôtel (hôtel); drop of the unstable "e": douch (douche) consonant contractions/clippings & abbreviations: dc (donc), pr (pour), ds (dans) ; double consonants: ele (elle), poura (pourra) ; semantic abbreviations (A.S): t (te/tu), p (peux/peut) agglutinations : jattends (j'attends)	9. 10. 11. a), 11. b) 11. c) 11. d), 11. e)
		12.

# suppression

graphical	typography & punctuation: [...] se genre de truc pr le site je pense ke ca devré allé vite je vou envéré [...] diacritic signs: ca (ça), voila (voilà)	13. 14.
-----------	--	------------

# addition

	repetition (characters, punctuation): suuuppeer !!!!	15.
graphical	mute character addition: peux (peu), as (a)	16.
	semiological representations (smileys/emoji) :-) 😊 💕 😊 😂 😊	17.
graphical	onomatopoeia: snif, bof	18.
phonetic	partial phonetic substitution: a) character addition, no phonetic modification: reparler (reparlé) b) liaisons: zaime (aime) phonetic substitution with variation: oki (ok), ouaip (oui)	19. 20.

## Examples for modified typology/Exemples correspondant à la typologie modifiée.

1. 30168, J ai **u** <PRE\_6> **o** tel hier soir (pdt que monsieur recevai...)elle est ravie...tu as assure sur ce coup la,@+

2. 3577, Cc <PRE\_1> !! ièr soir g pa insisté com ta pu le voir, **voulé pa ke** <SUR\_7> sénerv a koz 2 moi !!! 1é cour 2 robin d boi ce soir à 20 h, tjs **Dcidé a** yalé, si le **fé pa ce soir le feré jamé**. Pr l'épatatoes 2 vend., c tjs pa si on pt chanté «orevoir Présiden é J-<SUR\_4>» !! Kèl suspens !!! Ds le doute, préférable kon continu à **boC**. Repriz 2min (12 h) é jeudi matin pui stop pr 7 sem. T sur le piton ou tu te **bala2** ? Bon courage

3. 30657, Logiquement, si on s'embrasse, c'est que ça signifie quelque chose, mais dans la position où nous sommes je pense que la réponse est **nan** :)

4. 39406, Ouais mais je **m en** doutais à moitié, il **m en** avait déjà parlé. Le salaud il **m a** tjrs pas répondu!

5. 53770, Bon bah si déjà tu as trouvé ton plan ca va moi je trouve que c'est le + dur

6. 70343, Dit **mwa** pk tfai sa a **mwa jt** rien fait **mwa**

7. a) 71936, Ouais, faut que j'ramène à **tlm** des trucs --' mais j'sais pas quoi **lol**

7. b) 89810, Aaaaahhhh :p tu vient pas en cour demain ? Sinon **tfk** ? Et **qdb** ?

7. c) 92754, Oui! J'étais trop **deg!**

7. d) 31023, Regarde sur le **net**

7. e) 47607, Bouh α.α bon allez espère que ta flemme s'est arrangée un peu.. Un **zou\***

8. 5226, Vendredi promis. Je viendrai en scoot. Je **v** aller faire mon certificat.

9. 60286, Slt miss juste un **pétit** mot pour avoir de t news vu ke la dernière fois t'avais pas trop le moral. Bizoo

10. 43283, Pleins 2 bisous a vs 3. On pense a vs et on vs oublie pas. **just** 1 peu trop overbooké en ce moment. je **fini** le taf a 17h demain.j.essairai 2 vs tel.bisous a vs 3.2 ns 4.

11. a) 92836, Ben je vien **pr** 12h ... Dsl ...

11. b) 42632, 17 oct. 2011 11:03:39, 162,, Bon je pars de la fac la ^.^ **dc** a tte ^.^

11. c) 60840, 9 nov. 2011 11:06:52, 120,, Lol on **vera** dan kelk anè lol

11. d) 23631, tu **f koi**

11. e) 40789, **C** np koi **c** jeu lol

12. 44069, Oh tu me crois **jviens d'arriver jouvre** la porte mon portable libre :P ; 43315 : C'est bon, j'ai fini. **J't'attends** là bas. Bisous

13. 67645, Mais non c est pas une embrouille pour une histoire de carte grise qui a une solution mais cela ne change rien a votre amour elle t aime allez je te sers ton jaune je t attends on va passer une bonne soirée

14. 92 752, **A** chaud... Purre moi j'ai eu un **contrôle** sur stat, primitive et **étude** de fonction par rapport au **cout** marginal et d'autres truc comme **ca**, laisse tomber je me suis ramasser, alors que j'avais trop bosser, attends j'avais refait tt les exos j'avais tt bon et **la** au **contrôle** c'est que des truc hyper dur genre 2 fois plus dur qu'au **contrôle** ! Avec genre on voit pour **a la** puissance 2 et la c'était puissance 3 on savais pas comment faire pour l'exo ca faisait bizarre

15. 8427, Ahhh t'es **trooooooooooop foorte** !! :D :D j'y avais pas pensee mdrrrrrr

16. 32081, [...] Pour info, elle m'**as** dit qu'il y avait plusieurs filles qui voulait sortir avec lui, mais qu'il gardait ses distances. Si t'as besoins de précisions sur un passage, demande, je t'expliquerai mieux. Je te l'accorde, elle peut être un peu chiante....»...

17. 986, J viens de me faire virer par sms :(

35642 Et si on prenait le **bus** pour le **Gaumont** à **10** ? C'est que le **bus** commence à **10** et je ne sais pas combien de temps il fait pour y arriver et puis je pense qu'il y aura du monde.

18. 90944, Ben si j'rencontre personne **snif**

19. 53487, Non on en a pas reparler

4360, Cc <PRE\_1> ! Ça boum ? <PRE\_5> na k bien se tenir ! Oui, j sui **zalé** ! Pensé pa ke c t oçî fisik, v avoir lé **zépol** en Bton ! Ambiance tré 5pa. Si Papi ne cri pa o scandal, je continu. Pense pa mé fo ke lui dde 1 certif é me dira si c bon, en tt k pa mové, pr ce ke g ! 12 h à l'épad bien o cho é nouri ki + zé é ceriz sur le gato : av 2-2 : 2 koi je me plin ?!! Ta fé koi ièr soir ? 1 tètâtèt av le PC ? Bone journé. Biz

20. 86258, **Oki oki** Beh bon courage ^.^

### 10 tags

PRE (first name, 10,905)
SUR (nickname, 1,042)
NOM (last name, 785)
TEL (telephone number, 123)
LIE (place, 102)
ADR (address, 85)
MAR (brand name, 58)
COD (code, 50)
MEL (email address, 27)
URL (13)

Full corpus (88,000 French text messages)  
 2 samples (100 annotated sms, 1,000 transcoded sms)  
 available for free-of-charge download

› <http://88milsms.huma-num.fr>

### **anonymised sms** (n° 11326, 88milsMS corpus)

B, kèl intense réflexion ! Je c, t en week ! <SUR\_5> a A C 2 matièr pr fèr son suG. Concer tré 5pa ièr.  
 Bone soiré a toi é tte, bon week ? 2vé fèr gd bo ici : ke dal. Bisous.

### **anonymised & transcoded sms**

**Bon/Bien/Ben**, quelle intense réflexion ! Je sais, **tu es** en week-end ! <SUR\_5> a assez de matière pour faire son sujet. **Concert** très sympa hier. **Bonne soirée à toi et toute(s), bon week-end ? Il** de-vait faire grand beau ici : rien. Bisous.

### **Problems**

:B	semantic abbreviation
:t	oral/sociolinguists ('tes'/'tu es'), morpho-syntactic parsers ('tu t'es trompé')
:le	'ellipsis', not injected before 'concert', additional interpretation?
:Bone soiré...	ambiguity between 'to you' or 'see you later' (in French 'à toute')
:il	mandatory for morpho-syntactic parser processing
:que dalle	colloquial form, could be replaced by 'rien' ('nothing'), the standard form.

### **transcoding? 1,000 sms**

### **annotation? 100 sms**

difficulties for linguistic annotation, corresponding mainly to variations/modifications compared to standardised usage:

- › single or double (or more) tags?
- › most common double tag <MOD> & <ORT>; is spelling variation intentional?
- › if item is in *Petit Robert* dictionary, then a tag is not inserted
- › punctuation often reduced/absent in text messages, should it be introduced?

TYP	ography
MOD	ification
GRA	mmar
EMO	ji, smileys
ABS	cence, ellipsis
LAN	guage
ORT	ography/spelling
DIV	erse

### **SMS writing (eSMS)**

**highly variable writing forms:** aujourd'hui, ajoutd'hui, aujd, auji, aujii, aujiurd'hui, aujiurdhui, aujoirdhui, aujord'hui, aujordhui, aujordui, ajoutd'hui, Aujoud'hui, aujourd'hui, aujourdghuo, aujourdhhui, aujourdhui, aujourfui.

**consonant contractions/clippings:** slt, dsl

**apocope:** les appli sont pas encore a jour

**aphaeresis:** bon allez espère que ta flemme s'est arrangée un peu.. Un zou\*

**elision/agglutination:** Je c pa jtapel avant de sortir du gineco

**suppression of mute word-endings:** vou

**semantic abbreviations:** tu f koi ? (fais/feras/faisais/fous/foutais)

**more-or-less complex phonetic substitution:** koi, boC, 2m1

**repetitions/character addition:** suuuuppppeeerrrr, les zamours, oki,

**smileys/emoji ^^ :)**

**rich lexical creativity:** bisoutoucalinourienkepourtoipuissance, frontenormeetjouesdehamsterjovial