



**HAL**  
open science

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

Cédric Lopez, Mathieu Roche, Rachel Panckhurst

► **To cite this version:**

Cédric Lopez, Mathieu Roche, Rachel Panckhurst. Non-standard texts: from theoretical positions to Natural Language Processing normalisation. PLIN-Day, May 2016, Louvain-la-Neuve, Belgium. 2016. hal-01487025

**HAL Id: hal-01487025**

**<https://hal.science/hal-01487025>**

Submitted on 10 Mar 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Cédric Lopez\* Mathieu Roche\*\* Rachel Panckhurst\*\*\*



\*R&D, Viseo, Grenoble  
cedric.lopez@viseo.com

\*\*UMR TETIS, Cirad, Irstea, AgroParisTech, Montpellier;  
LIRMM, UMR 5506 CNRS & Université Montpellier  
mathieu.roche@cirad.fr

\*\*\*Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3  
rachel.panckhurst@univ-montp3.fr

## 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

88milSMS corpus

## 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

knowledge building

SMS writing (eSMS)

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

language classification

C1.1: LEFF  
C1.2: LEFF no accents

C2.1 (sole letters) a, c, f, j, p, v...

C2.2 (time) 8heures, 10minaperdre, 6-7h

C2.3 (repetition) Mdrrr, Lool, tkkkkt, HUUUmmm

C2.4 (special) Conn\*rd, désannule, resto+ciné

C2.5 (numbers) numb3rs, mc2, 106ounette, 3615ma-vie

C2.6 (smileys) ^^ :p ;) :d <3 :-) xd :( :/

C3 (INSO)

tkt, jte, cc, voituration, cinglicité, tetrangle

## 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)