AGORA

iut Cergy-Pontoise

UNIVERSITÉ de Cergy-Pontoise

"A Textometrical Analysis of French Arts Workers “fr.*Intermittents*” on Twitter

J LONGHI & D SAIGH

IRD2016

R2DIP
Réseau de Recherches des Discours Institutionnels & Politiques

CORPUSCOMERE
COMMUNICATION MÉDIÉE PAR LES RÉSEAUX

Plan

- Definitions & Related projects
- Introduction
- Corpus #Intermittent: context and constitution
- First analysis with Iramuteq & Lexico 3
- Future work
- Conclusion

Related projects

- Transdisciplinary projects
  - “**Digital Humanities and Data Journalism**” (2014)
  - **ICI 2.0 (2015)**
  - **#Idéo2017**
- bring together researchers in the fields of linguistic research (discourse analysis) and computer science. Create tools or applications

- “**CoMeRe**” project: create a core corpus of Computer-Mediated communications (CMC) in French . For different communication systems: blogs, tweets, SMS / texting, emails, forums, etc …

INTRODUCTION (1)

- Using social media for social discourse is becoming common practice to express the different ideologies/point of view

- Twitter : “new genre of discourse” (Longhi, 2013)

- Twitter offers the opportunity to citizens to express themselves concisely but quickly and with less preparation exposing them easier to public

- Because of its brevity and clarity: reflects a **semantic condensation** which can be favorable to the **ideologies and controversies**

INTRODUCTION (2)

Analysis of social networks:

- Semantic technologies: data on social networks annotated /indexed by ontologies

- Semantic analysis of online social networks: semantics are often only considered; the symbolic and semiotic dimensions of textual data are not used efficiently.

→ **Use linguistic techniques dedicated to discourse analysis for NLP**

- The challenge is to keep the criteria and methods of processing natural language texts, but with research from discourse analysis

Discourse analysis

- Part of social sciences
- Multidisciplinary field (history, politics, media, literature, education, etc.).
- Considering the structure of a text by relating it to its conditions of production, let's understand it as a **discourse**.
- There are various approaches to discourse analysis: textual analysis is the most often used form in which discourse is analysed.
- It is through discourse that political ideologies are acquired, expressed, learned, propagated and contested (Van Dijk, 2006)

Textometry

- Textometry, born in France in the 80's, has developed powerful techniques for the analysis of large body of texts. Following lexicometry and text statistical analysis, it offers tools and methods tested in multiple branches of the humanities and is statistically well founded. http://textometrie.ens-lyon.fr/?lang=en

- Textometry is particularly relevant to corpus exploitation in human and social sciences.
- It simultaneously enables a detailed and global observation of different texts while remaining close to them, and highlights the fact that language is an important observation field for human and social sciences.

Corpus #intermittent: The ‘new’ agreement

In March 2014, social partners signed a new agreement concerning the unemployment benefits for French arts workers. This text that became the convention of 14 May 2014 on unemployment benefits aroused concerns and opposition among the arts workers.

A protest movement and mass demonstrations took place in Paris and in other French cities and lasted for several days.

These reactions rapidly invaded social networks especially Twitter. Millions of tweets were written as soon as the first information about this controversy emerged.

Goal of the corpus

obtain a corpus which enables us to work on this kind of discourse (tweets related to a controversial topic)

characterize it and understand in order to extend previous research (Longhi 2006, 2008) that focused on French arts workers in 2003/2004.

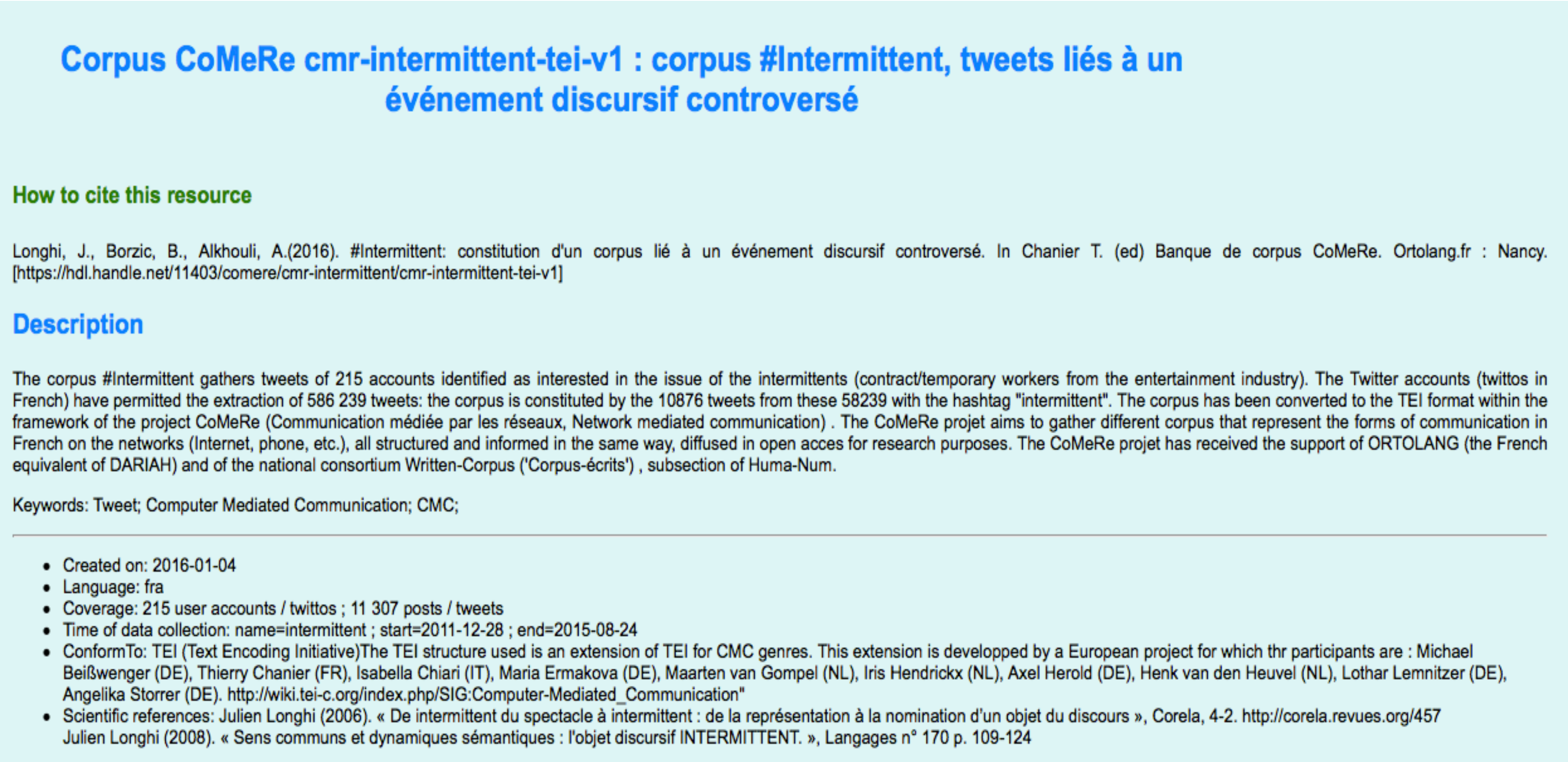
Process

In 2014: 13 074 tweets with *#intermittent* posted by 4 617 people.

In 2015: we established a threshold of at least 10 tweets with *#intermittent*: we obtained 215 accounts that had produced at least 10 tweets explicitly referenced as belonging to this theme (in order to have representative accounts).

By collecting all the tweets from these 215 people, we gathered 586 239 tweets that included 10 876 tweets with *#intermittent*. The corpus *#intermittent* corresponds to these 10 876 tweets.

#Intermittent



Corpus CoMeRe cmr-intermittent-tei-v1 : corpus #Intermittent, tweets liés à un événement discursif controversé

How to cite this resource

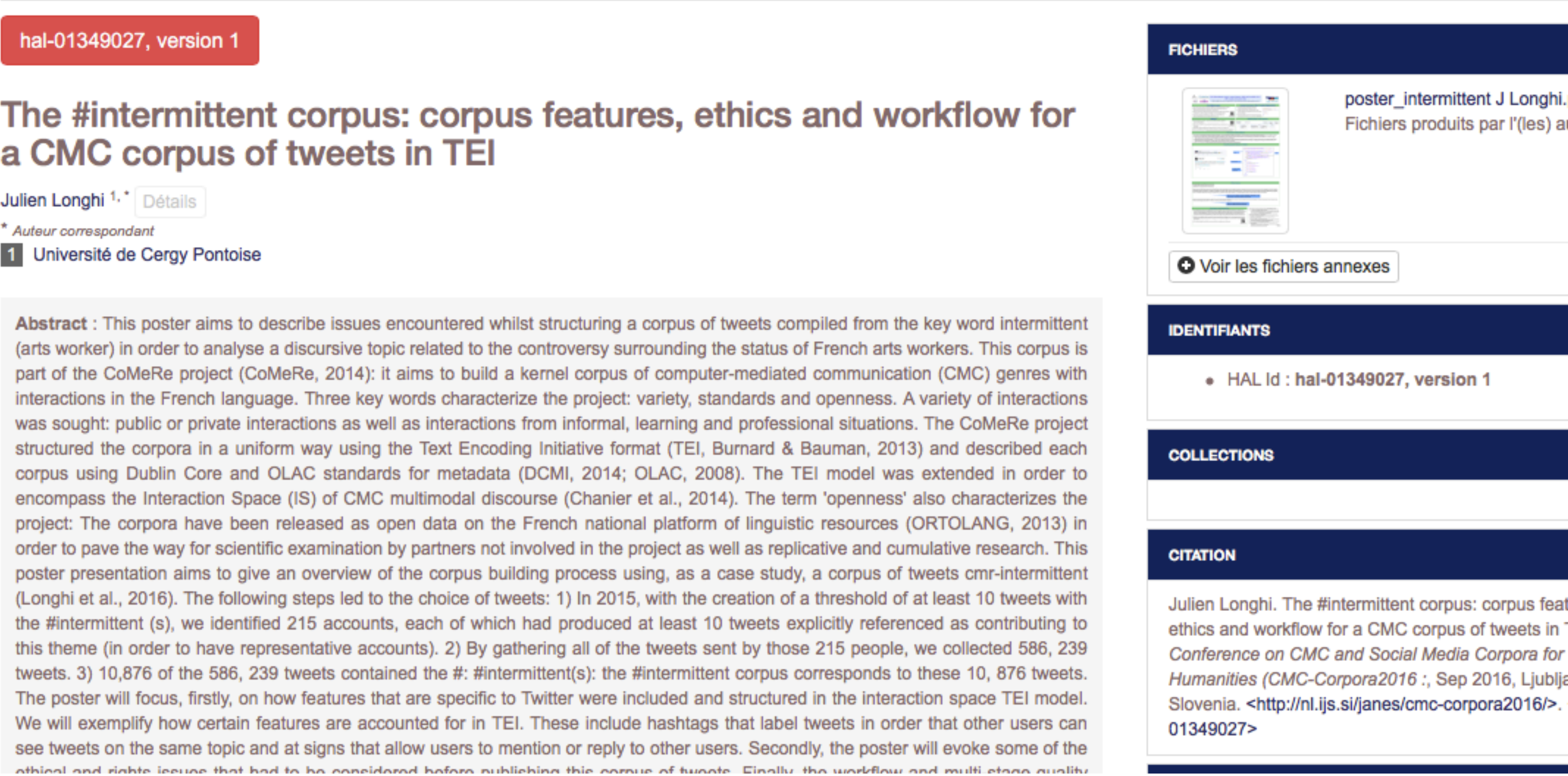
Longhi, J., Borzic, B., Alkhouli, A.(2016). #Intermittent: constitution d'un corpus lié à un événement discursif controversé. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [https://hdl.handle.net/11403/comere/cmr-intermittent/cmr-intermittent-tei-v1]

Description

The corpus #Intermittent gathers tweets of 215 accounts identified as interested in the issue of the intermittents (contract/temporary workers from the entertainment industry). The Twitter accounts (twittos in French) have permitted the extraction of 586 239 tweets: the corpus is constituted by the 10876 tweets from these 58239 with the hashtag "intermittent". The corpus has been converted to the TEI format within the framework of the project CoMeRe (Communication médiée par les réseaux, Network mediated communication) . The CoMeRe projet aims to gather different corpus that represent the forms of communication in French on the networks (Internet, phone, etc.), all structured and informed in the same way, diffused in open acces for research purposes. The CoMeRe projet has received the support of ORTOLANG (the French equivalent of DARIAH) and of the national consortium Written-Corpus ('Corpus-écrits') , subsection of Huma-Num.

Keywords: Tweet; Computer Mediated Communication; CMC;

- Created on: 2016-01-04
- Language: fra
- Coverage: 215 user accounts / twittos ; 11 307 posts / tweets
- Time of data collection: name=intermittent ; start=2011-12-28 ; end=2015-08-24
- ConformTo: TEI (Text Encoding Initiative)The TEI structure used is an extension of TEI for CMC genres. This extension is developped by a European project for which thr participants are : Michael Beißwenger (DE), Thierry Chanier (FR), Isabella Chiari (IT), Maria Ermakova (DE), Maarten van Gompel (NL), Iris Hendrickx (NL), Axel Herold (DE), Henk van den Heuvel (NL), Lothar Lemnitzer (DE), Angelika Storrer (DE). http://wiki.tei-c.org/index.php/SIG:Computer-Mediated_Communication"
- Scientific references: Julien Longhi (2006). « De intermittent du spectacle à intermittent : de la représentation à la nomination d'un objet du discours », Corela, 4-2. http://corela.revues.org/457
  Julien Longhi (2008). « Sens communs et dynamiques sémantiques : l'objet discursif INTERMITTENT. », Langages n° 170 p. 109-124



hal-01349027, version 1

The #intermittent corpus: corpus features, ethics and workflow for a CMC corpus of tweets in TEI

Julien Longhi [1], * Détails

* *Auteur correspondant*

1 Université de Cergy Pontoise

Abstract : This poster aims to describe issues encountered whilst structuring a corpus of tweets compiled from the key word intermittent (arts worker) in order to analyse a discursive topic related to the controversy surrounding the status of French arts workers. This corpus is part of the CoMeRe project (CoMeRe, 2014): it aims to build a kernel corpus of computer-mediated communication (CMC) genres with interactions in the French language. Three key words characterize the project: variety, standards and openness. A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations. The CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and described each corpus using Dublin Core and OLAC standards for metadata (DCMI, 2014; OLAC, 2008). The TEI model was extended in order to encompass the Interaction Space (IS) of CMC multimodal discourse (Chanier et al., 2014). The term 'openness' also characterizes the project: The corpora have been released as open data on the French national platform of linguistic resources (ORTOLANG, 2013) in order to pave the way for scientific examination by partners not involved in the project as well as replicative and cumulative research. This poster presentation aims to give an overview of the corpus building process using, as a case study, a corpus of tweets cmr-intermittent (Longhi et al., 2016). The following steps led to the choice of tweets: 1) In 2015, with the creation of a threshold of at least 10 tweets with the #intermittent (s), we identified 215 accounts, each of which had produced at least 10 tweets explicitly referenced as contributing to this theme (in order to have representative accounts). 2) By gathering all of the tweets sent by those 215 people, we collected 586, 239 tweets. 3) 10,876 of the 586, 239 tweets contained the #: #intermittent(s): the #intermittent corpus corresponds to these 10, 876 tweets. The poster will focus, firstly, on how features that are specific to Twitter were included and structured in the interaction space TEI model. We will exemplify how certain features are accounted for in TEI. These include hashtags that label tweets in order that other users can see tweets on the same topic and at signs that allow users to mention or reply to other users. Secondly, the poster will evoke some of the ethical and rights issues that had to be considered before publishing this corpus of tweets. Finally, the workflow and multi-stage quality

FICHIERS

poster_intermittent J Longhi.

Fichiers produits par l'(les) au

Voir les fichiers annexes

IDENTIFIANTS

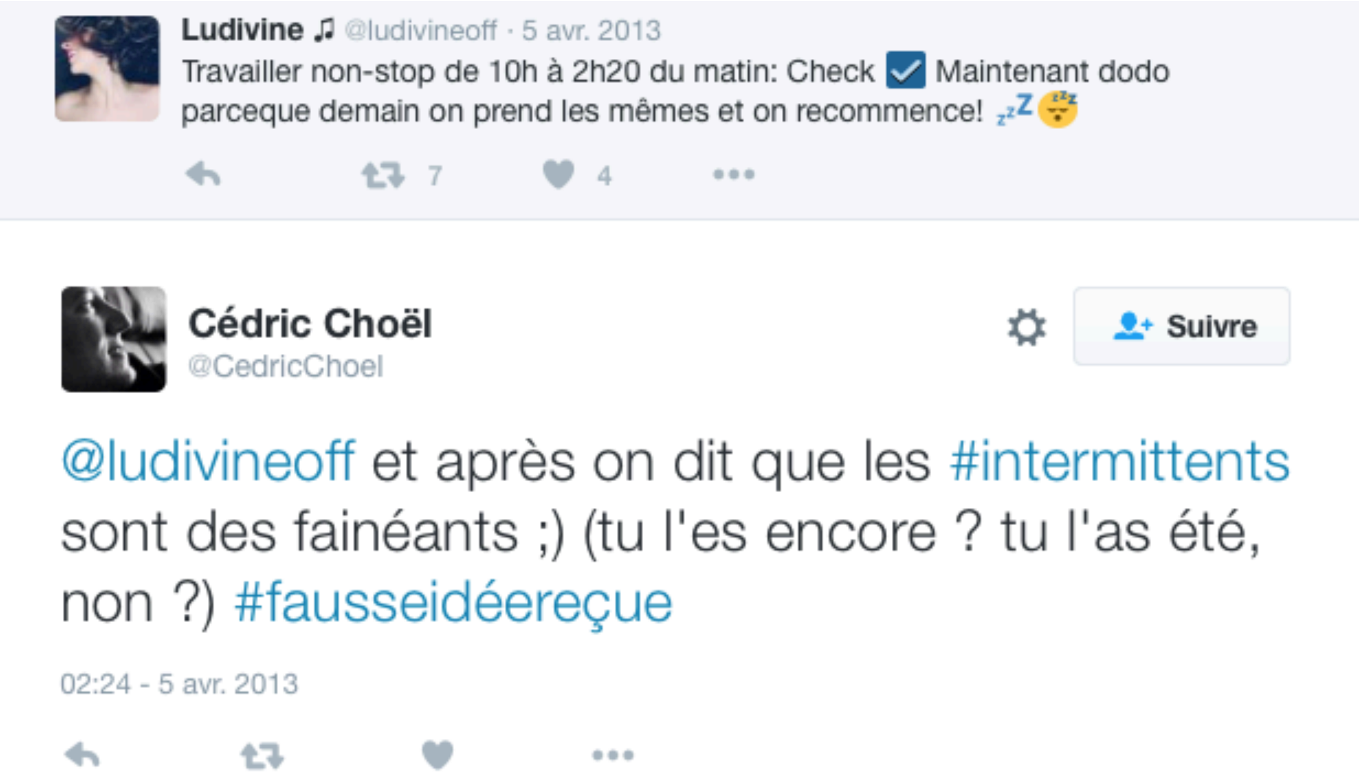
- HAL Id : **hal-01349027, version 1**

COLLECTIONS

CITATION

Julien Longhi. The #intermittent corpus: corpus feat ethics and workflow for a CMC corpus of tweets in T *Conference on CMC and Social Media Corpora for Humanities (CMC-Corpora2016* :, Sep 2016, Ljublja Slovenia. <http://nl.ijs.si/janes/cmc-corpora2016/>. 01349027>

Example



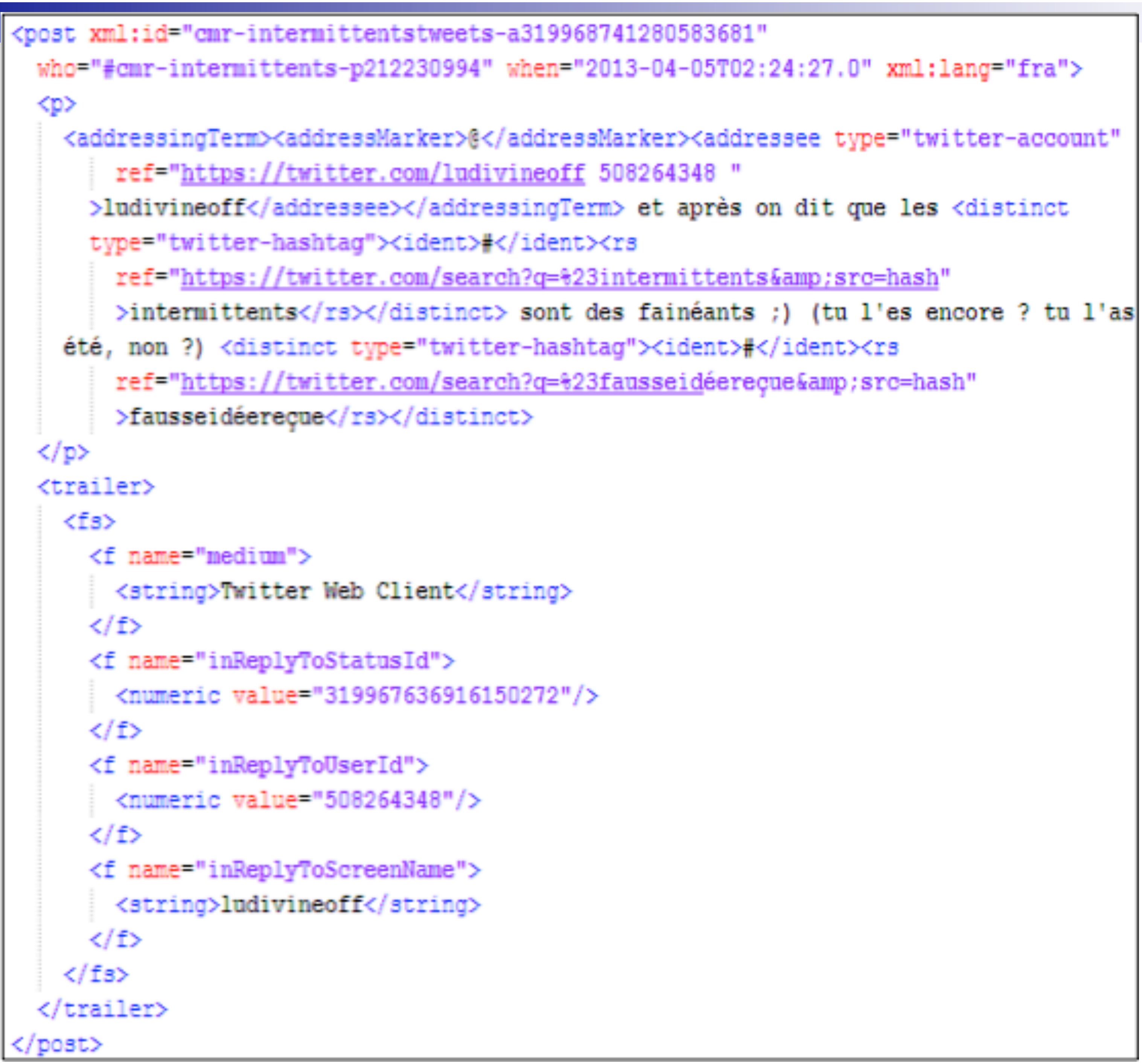
Ludivine ♫ @ludivineoff · 5 avr. 2013

Travailler non-stop de 10h à 2h20 du matin: Check ☑ Maintenant dodo parceque demain on prend les mêmes et on recommence! 💤😴

Cédric Choël
@CedricChoel

@ludivineoff et après on dit que les #intermittents sont des fainéants ;) (tu l'es encore ? tu l'as été, non ?) #fausseidéereçue

02:24 - 5 avr. 2013



```
<post xml:id="cmr-intermittentstweets-a319968741280583681"
  who="#cmr-intermittents-p212230994" when="2013-04-05T02:24:27.0" xml:lang="fra">
  <p>
    <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
        ref="https://twitter.com/ludivineoff 508264348 "
      >ludivineoff</addressee></addressingTerm> et après on dit que les <distinct
      type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23intermittents&amp;src=hash"
        >intermittents</rs></distinct> sont des fainéants ;) (tu l'es encore ? tu l'as
    été, non ?) <distinct type="twitter-hashtag"><ident>#</ident><rs
        ref="https://twitter.com/search?q=%23fausseidéereçue&amp;src=hash"
        >fausseidéereçue</rs></distinct>
  </p>
  <trailer>
    <fs>
      <f name="medium">
        <string>Twitter Web Client</string>
      </f>
      <f name="inReplyToStatusId">
        <numeric value="319967636916150272"/>
      </f>
      <f name="inReplyToUserId">
        <numeric value="508264348"/>
      </f>
      <f name="inReplyToScreenName">
        <string>ludivineoff</string>
      </f>
    </fs>
  </trailer>
</post>
```

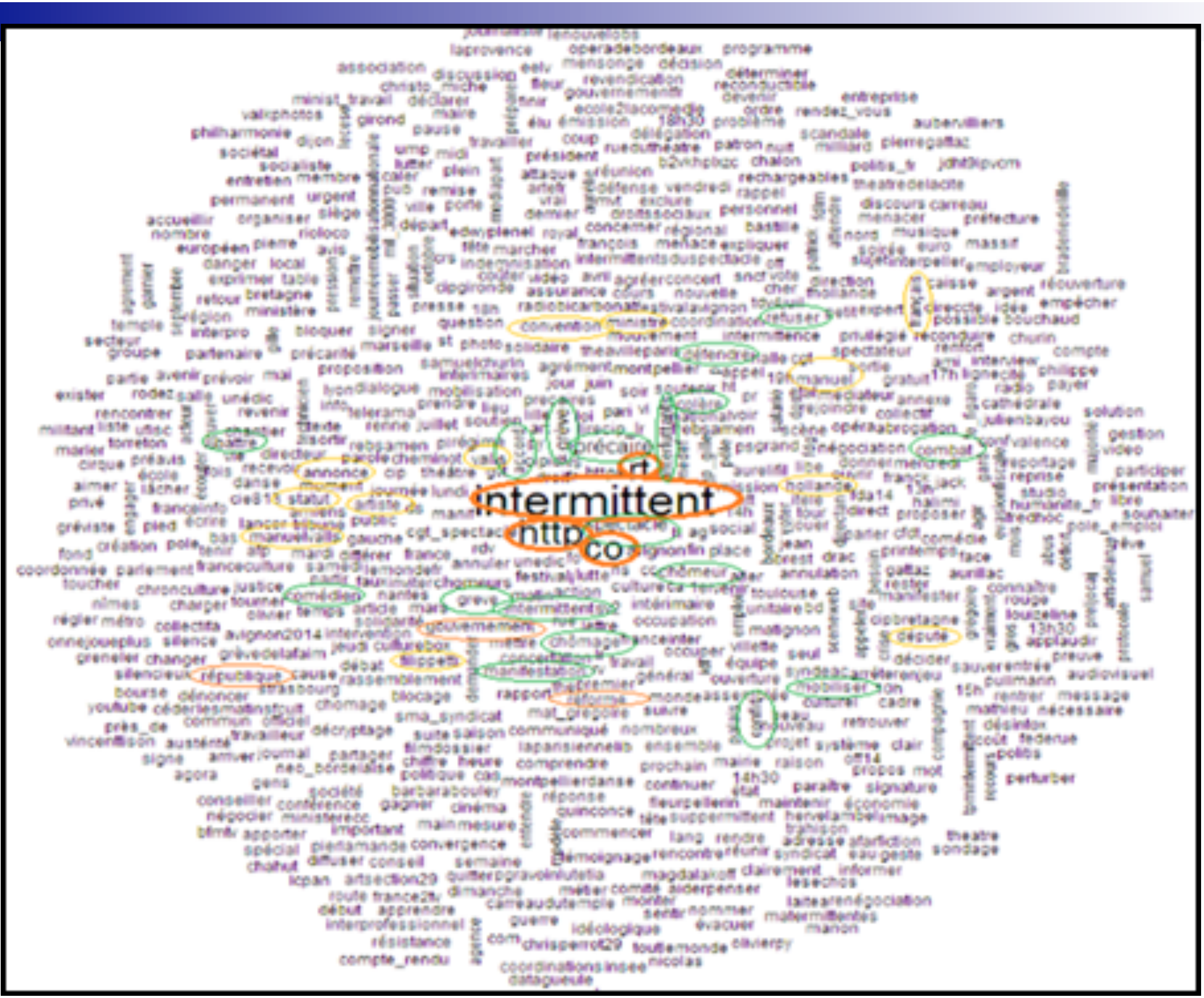
Iramuteq

- The Iramuteq software offers a set of analysis procedures for the description of a textual corpus.

- One of its principal methods is Alceste. This allows a user to segment a corpus into “context units”, to make comparisons and groupings of the segmented corpus according to the lexemes contained within it, and then to seek “stable distributions” (Reinert, 1998).

- In addition to the Alceste method, Iramuteq provides other analysis tools including prototypical analysis, similarities analysis, and word clouds analysis.

- All of these methods allow the users of this tool to map out the dynamics of the discourses of the different subjects engaged in interaction (Reinert, 1999).


intermittent
http
co
rt
convention
ministre
refuser
défendre
manuel
colère
précaire
combat
hollande
annonce
statut
artiste
valls
manuelvalls
comédien
grève
intermittentsp
gouvernement
chômage
chômeur
manifestation
république
filippetti
député
mobiliser
réforme
spectacle
laprovence
operadebordeaux
programme
association
discussion
mensonge
décision
déterminer
reconductible
entreprise
philharmonie
sociétal
socialiste
entretien
membre
permanent
urgent
accueillir
nombre
européen
danger
exprimer
bretagne
retour
ministère
région
interpro
temple
secteur
groupe
partenaire
partie
avenir
prévoir
exister
rencontrer
militant
marier
cirque
aimer
privé
gréviste
fond
coordonnée
parlement
toucher
nîmes
régler
métro
collectifs
silence
grenelle
changer
silencieux
bourse
dénoncer
youtube
vincent
signe
agora
conseiller
conférence
négocier
apporter
spécial
diffuser
conseil
semaine
important
convergence
résistance
compte_rendu
coordinations
insee
nicolas
festival
lutte
culture
toulouse
nantes
avignon
occupation
intérimaire
occuper
villette
seul
syndicat
cadre
retrouver
système
clair
prochain
raison
propos
mot
signature
économie
continuer
paraître
sondage
théâtre
adresse
informer
solution
gestion
vidéo
participer
présentation
libre
souhaiter
pôle_emploi
rouge
applaudir
preuve
audiovisuel
message
nécessaire
pullmann
fédérer
perturber
réouverture
argent
empêcher
bouchaud
compte
philippe
payer
radio
cathédrale
julienbayou
réforme
direction
caisse
employeur
massif
préfecture
musique
discours
menacer
politis_fr
theatredelacite
aubervilliers
pierregattaz
scandale
milliard
rendez_vous
programme
paris
marseille
montpellier
lyon
rennes
bordeaux
samedi
lundi
jeudi
juillet
juin
france
gauche
public
danse
école
théâtre
cinéma
sma_syndical
communiqué
travailleur
journal
partager
chômage
blocage
rassemblement
débat
cause
strasbourg
général
ouverture
équipe
travail
rapport
premier
suivre
nombreux
comprendre
ensemble
réponse
commencer
témoignage
rencontre
réunir
gaffe
diviser
apprendre
début
interprofessionnel
guerre
idéologique
évacuer
sénat
nommer
comité
penser
métier
dimanche
important
semaine

The word-cloud

This word-cloud highlights the most common occurrences in tweets. These lexical items are positioned centrally in the cloud. The occurrence "*intermittent*" is the largest in size because it constitutes the key word of our corpus; this is why its frequency is higher.

That word is followed by specific markers such as "co" and "http" that refer to links shared on Twitter (links automatically abbreviated).

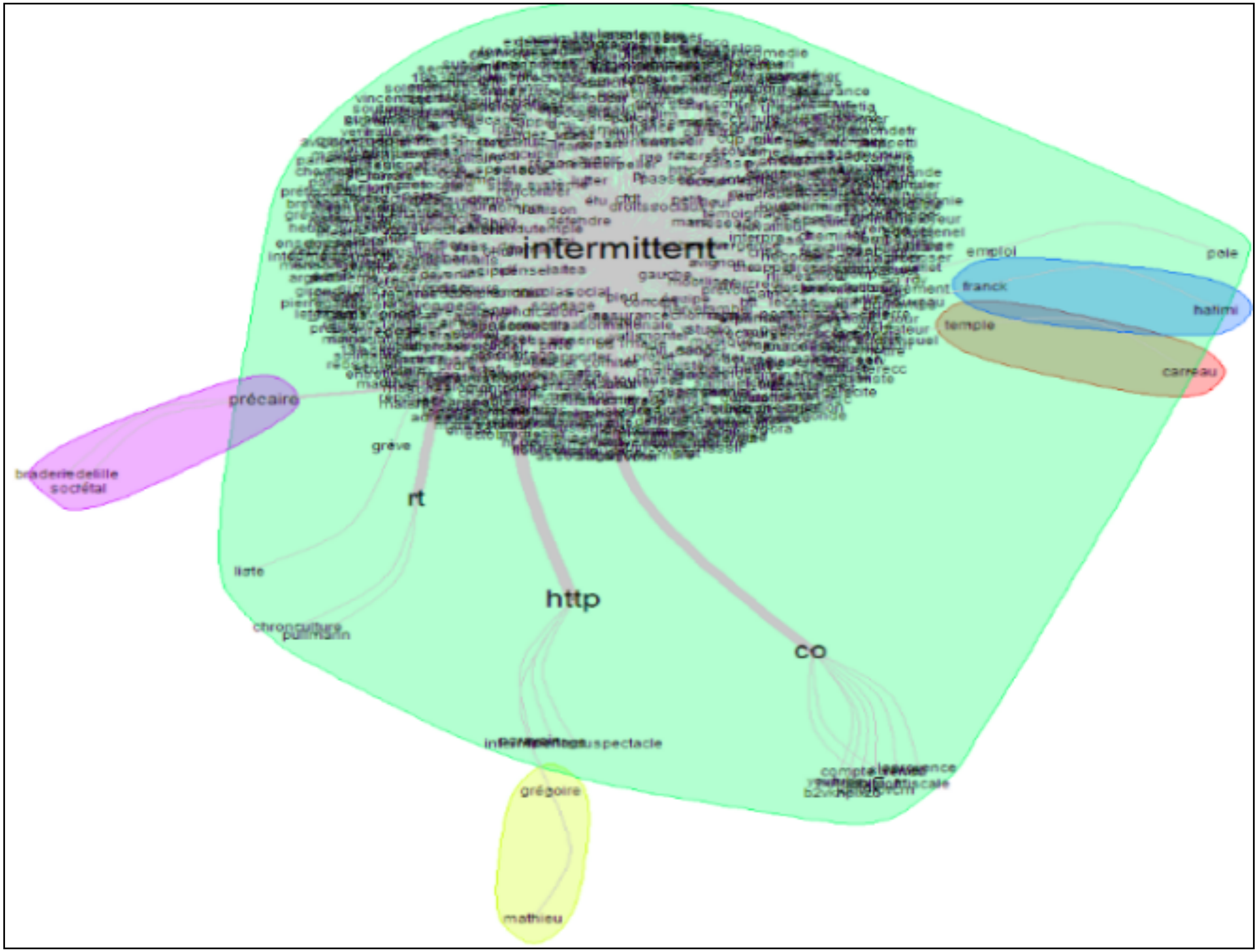
There is also the sign "rt" which means "retweet". This has the function of reposting the tweet of another person enabling users to quickly share it with all subscribers.

Similarities Analysis

Similarities analysis is a technique based on graph theory (Flament, 1962).

It presents in a graphical format the structure of a corpus, distinguishing between the shared parts and the specificities of coded variables.

This allows the link between the different forms in the text segments to emerge (Marchand & Ratinaud, 2012).


intermittent
défendre
gauche
avignon
assurance
emploi
pole
franck
hatimi
temple
carreau
précaire
braderiedelille
sociétal
grève
rt
liste
chronoculture
pullmann
http
intermittentsduspectacle
grégoire
mathieu
co

Similarities Analysis

The first observation that we can make is that this corpus is very homogeneous with one central idea around which revolves the greatest part of the lexicon of our corpus.

This figure shows a single main cluster, with some others which are very small and not relevant. This cluster consists of a word cloud which contains the key word "*intermittent*" at its center and around it, are grouped a very dense and related lexicon.

http

“*http”* group in which we find the term *intermittentdespectacle* (arts workers)

a little further, a small cluster containing the name *Gregory Mathieu*, a sociologist who wrote a book with the title “*Les intermittents du spectacle. Enjeux d’un siècle de luttes”*.

So, in these groups, we understand that the majority of links mentioned in tweets refer users to web pages where the name of the sociologist is mentioned.

rt

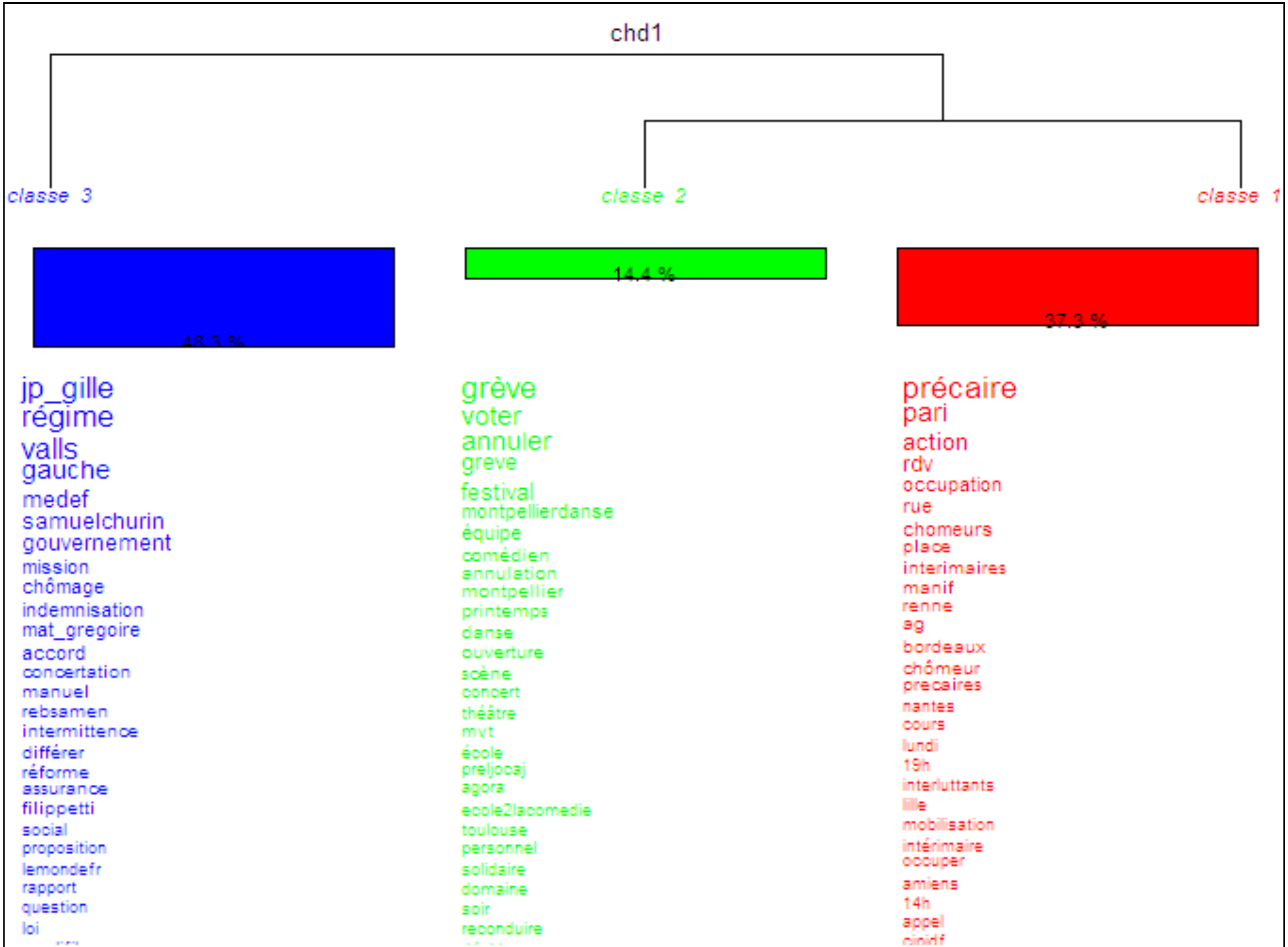
There is also the “*rt*” group which includes the following terms: *chronculture, pullmarin, dinamopress, angelin*... which refer to the names of accounts which have retweeted the most. The "co" group is, as explained above, the abbreviated form of links on Twitter.

We can already understand from this figure that the *#intermittent* corpus contains a lot of links, retweets related to French arts workers, and it describes their various actions and their status (highlighted by the cluster *précaires* (precarious).

The Hierarchical Descending Classification

One method used by Alceste is the hierarchical descending classification. This method offers a global approach to a corpus.

The *HDC* after partitioning the corpus, identifies statistically independent word classes (forms). These classes are interpreted through their profiles, which are characterized by specific correlated forms. The *HDC* shows that using a dendrogram.



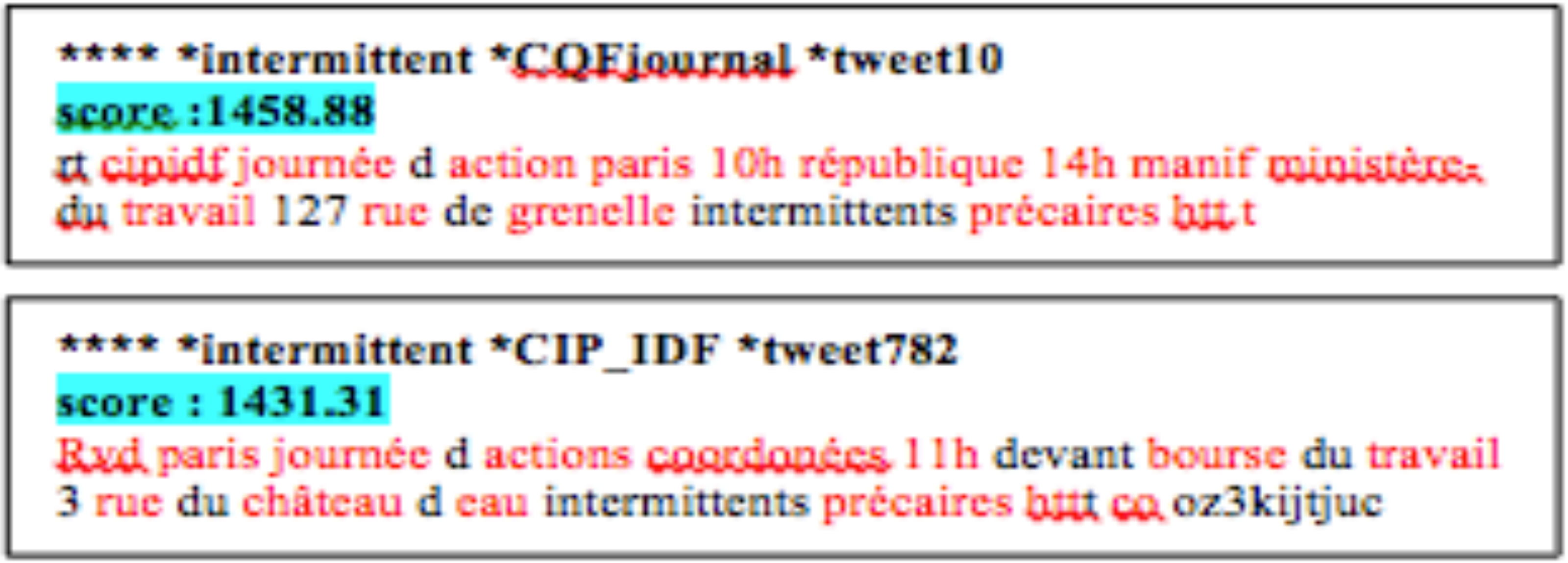
Dendrogramme CHD1 - phylogram

Classes

Two groups are distinguished in this figure, the first with two related classes (class 1 and class 2), and the second where there is only one class (class 3).

The class 1 includes forms associated with the different protest movements of French arts workers such as the occupation of streets, theaters and other places, the demonstrations in Paris and elsewhere.

Class 1



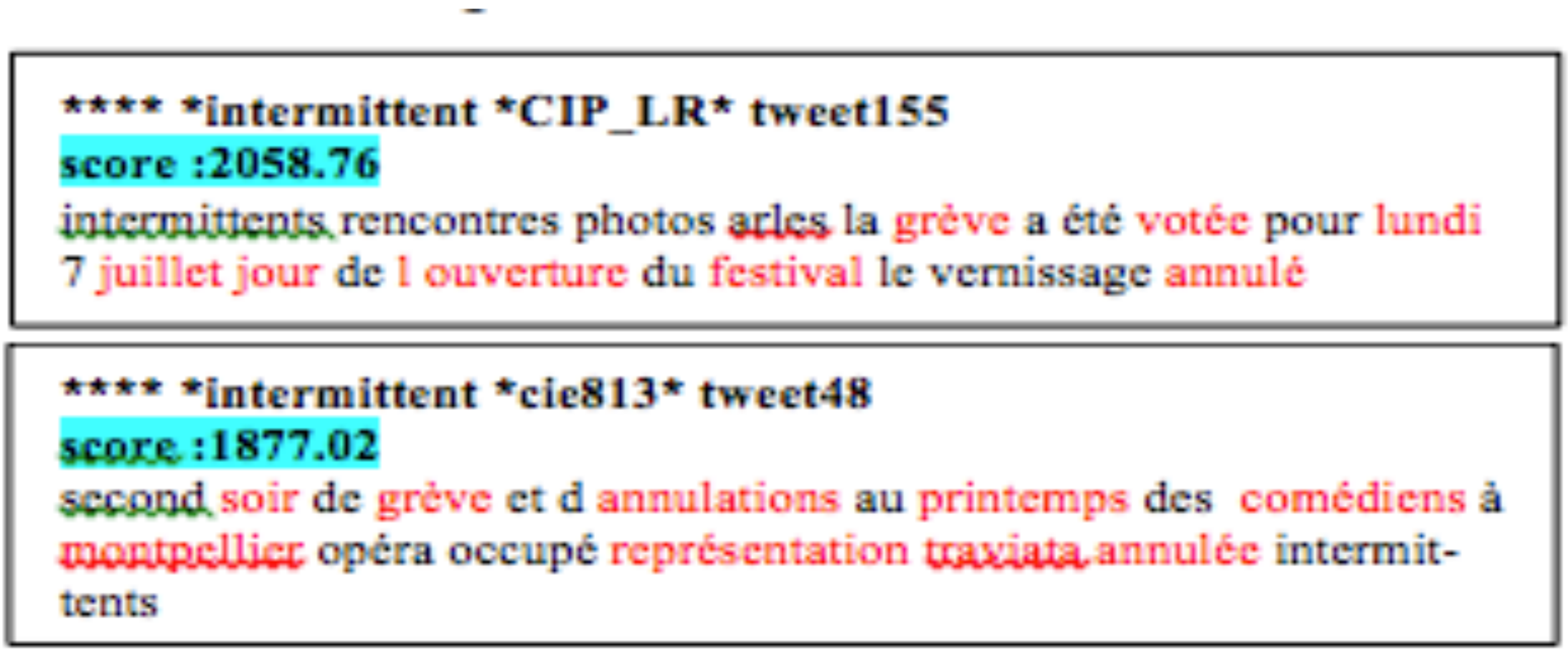
**** *intermittent *CQFjournal *tweet10
score :1458.88
rt cipidf journée d action paris 10h république 14h manif ministère du travail 127 rue de grenelle intermittents précaires htt t

**** *intermittent *CIP_IDF *tweet782
score : 1431.31
Rvd paris journée d actions coordonées 11h devant bourse du travail 3 rue du château d eau intermittents précaires http co oz3kijtjuc

Classe 2

Class 2 refers to strikes held by the French arts workers and their different concerts and show cancellations. This class contains words such as: *grève* (strike), *festival* (festival), *annulé* (cancelled).

Class 2



**** *intermittent *CIP_LR* tweet155

score :2058.76

intermittents rencontres photos arles la grève a été votée pour lundi 7 juillet jour de l ouverture du festival le vernissage annulé

**** *intermittent *cie813* tweet48

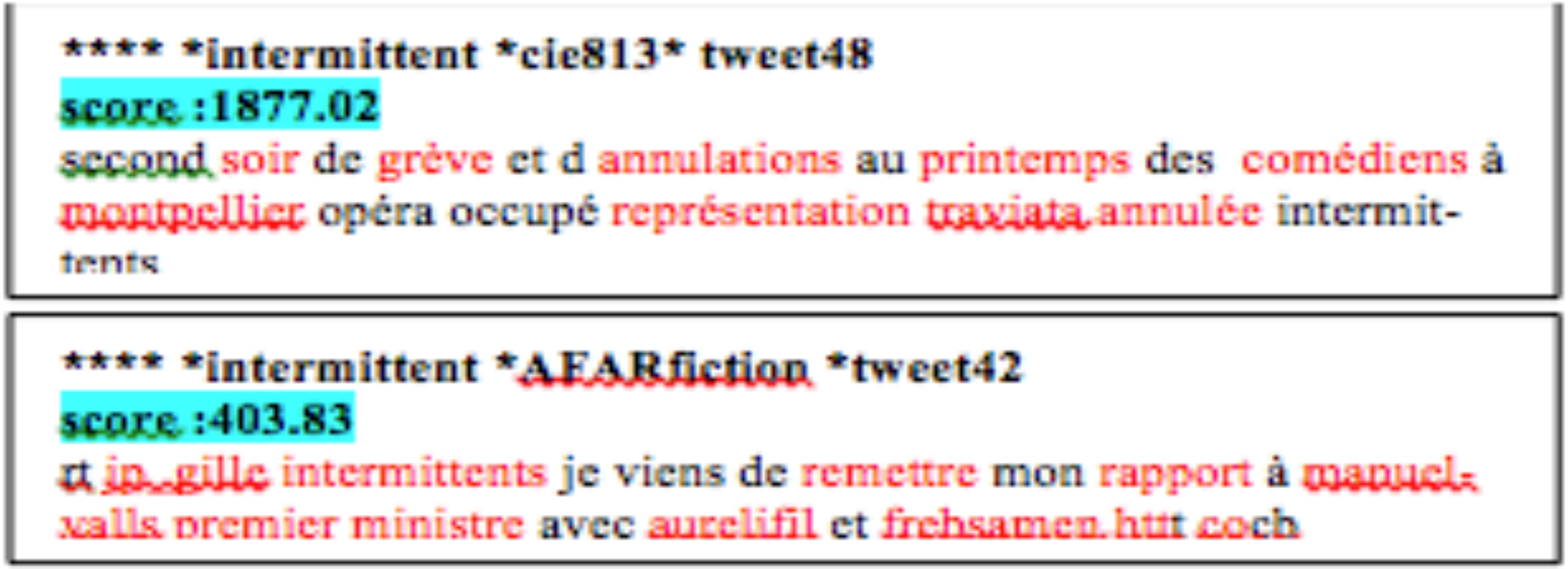
score :1877.02

second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

Classe 3

Class 3 concerns the tweets that talk about the unemployment insurance system related to the French arts workers and political entities involved in this affair. Here is a characteristics segment summarizing the words associated with this class, including *medef*, *valls*, *samuelchurin*, *aurelifi*

Class 3



**** *intermittent *cie813* tweet48

score :1877.02

second soir de grève et d annulations au printemps des comédiens à montpellier opéra occupé représentation traviata annulée intermittents

**** *intermittent *AFARfiction *tweet42

score :403.83

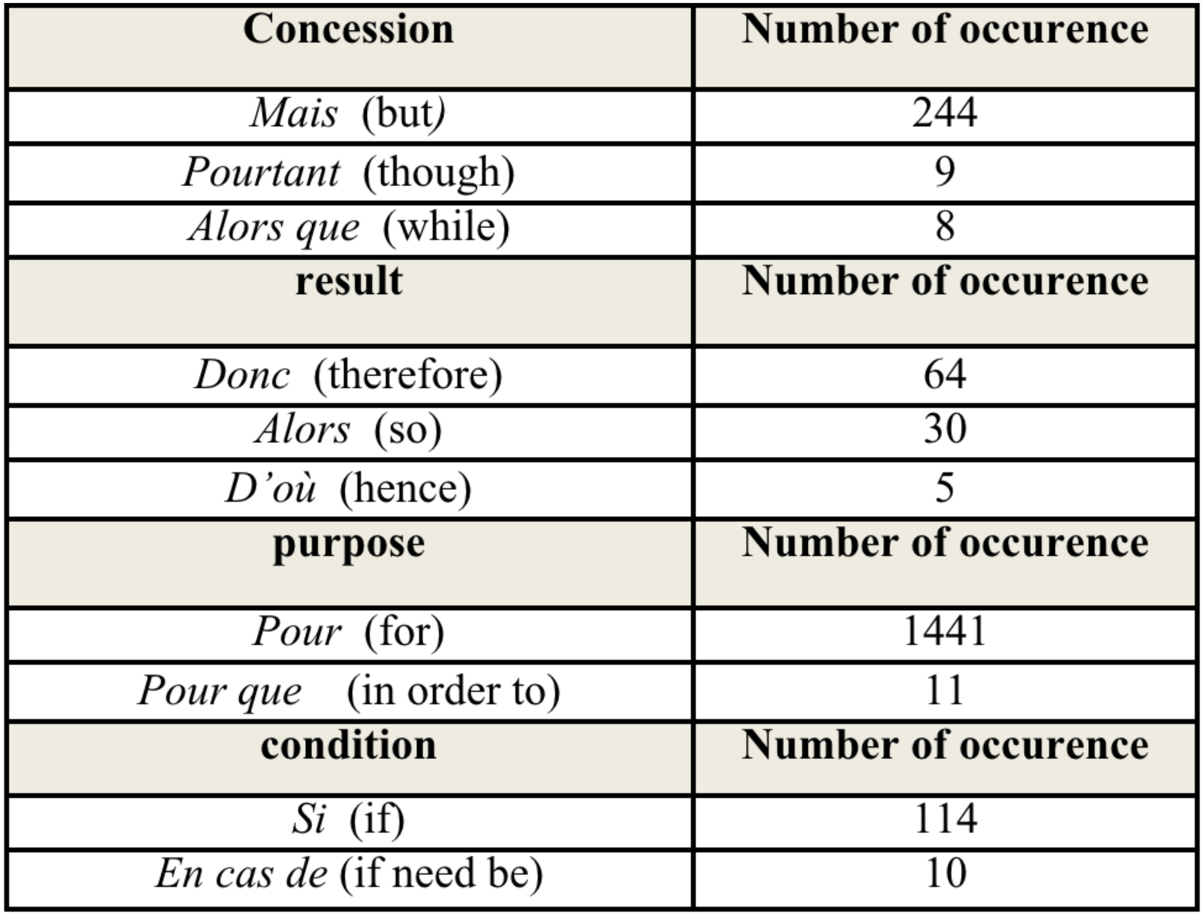
rt jp_gille intermittents je viens de remettre mon rapport à manuel-valls premier ministre avec aurelifil et frebsamen http tcoch

A more discursive analysis with Lexico 3

Lexico is textual data analysis software.

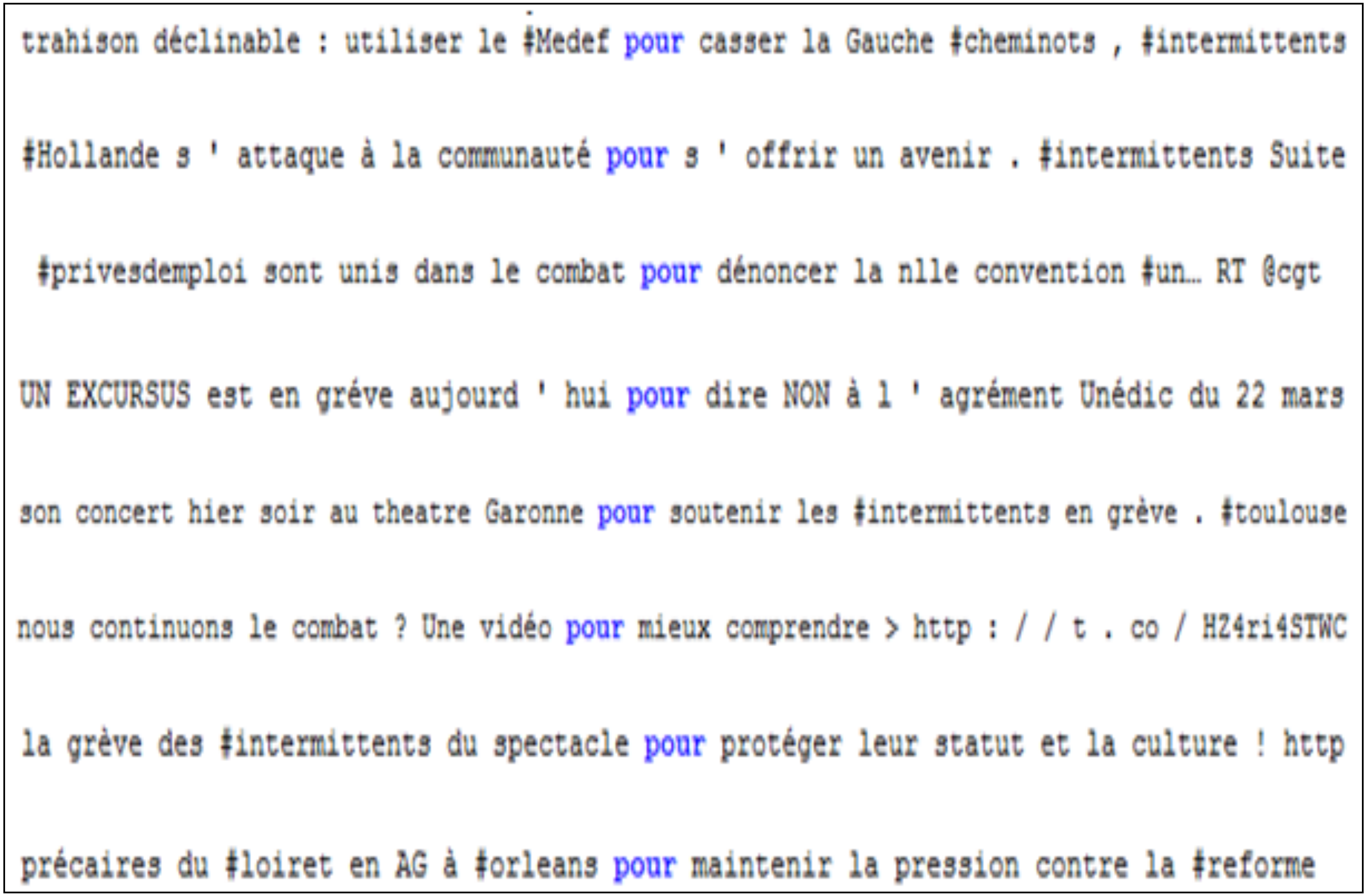
Develloped by the statistician Andrew Salem and his team, this software allows exploring a corpus of texts through the vocabulary, then comparing this corpus, previously segmented into parts, according to the vocabulary of each one.

Connectors



Concession	**Number of occurence**
Mais (but)	244
Pourtant (though)	9
Alors que (while)	8
result	**Number of occurence**
Donc (therefore)	64
Alors (so)	30
D'où (hence)	5
purpose	**Number of occurence**
Pour (for)	1441
Pour que (in order to)	11
condition	**Number of occurence**
Si (if)	114
En cas de (if need be)	10

Pour: goal



trahison déclinable : utiliser le #Medef pour casser la Gauche #cheminots , #intermittents

#Hollande s ' attaque à la communauté pour s ' offrir un avenir . #intermittents Suite

#privesdemploi sont unis dans le combat pour dénoncer la nlle convention #un… RT @cgt

UN EXCURSUS est en gréve aujourd ' hui pour dire NON à l ' agrément Unédic du 22 mars

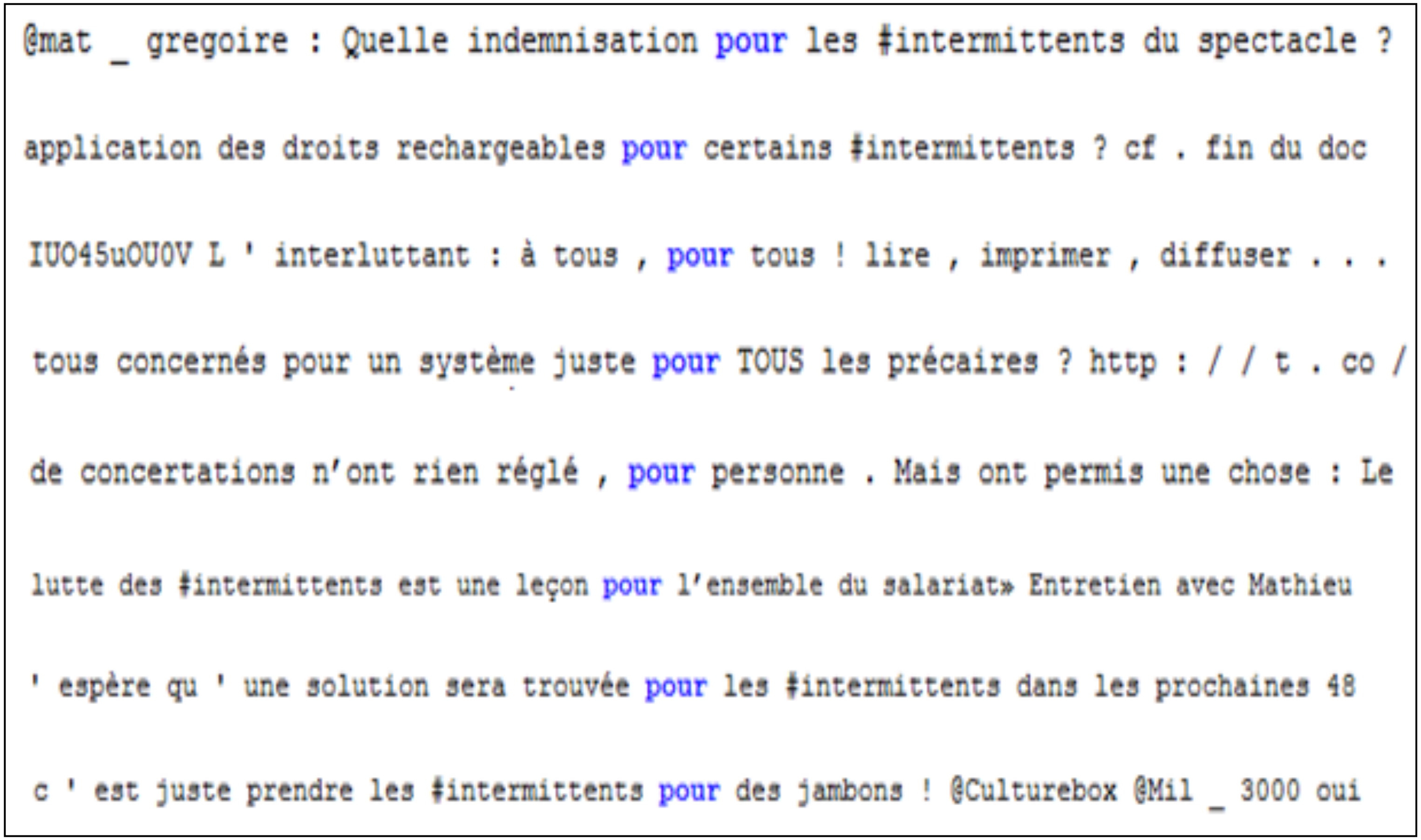
son concert hier soir au theatre Garonne pour soutenir les #intermittents en grève . #toulouse

nous continuons le combat ? Une vidéo pour mieux comprendre > http : / / t . co / HZ4ri4STWC

la grève des #intermittents du spectacle pour protéger leur statut et la culture ! http

précaires du #loiret en AG à #orleans pour maintenir la pression contre la #reforme

Pour préposition



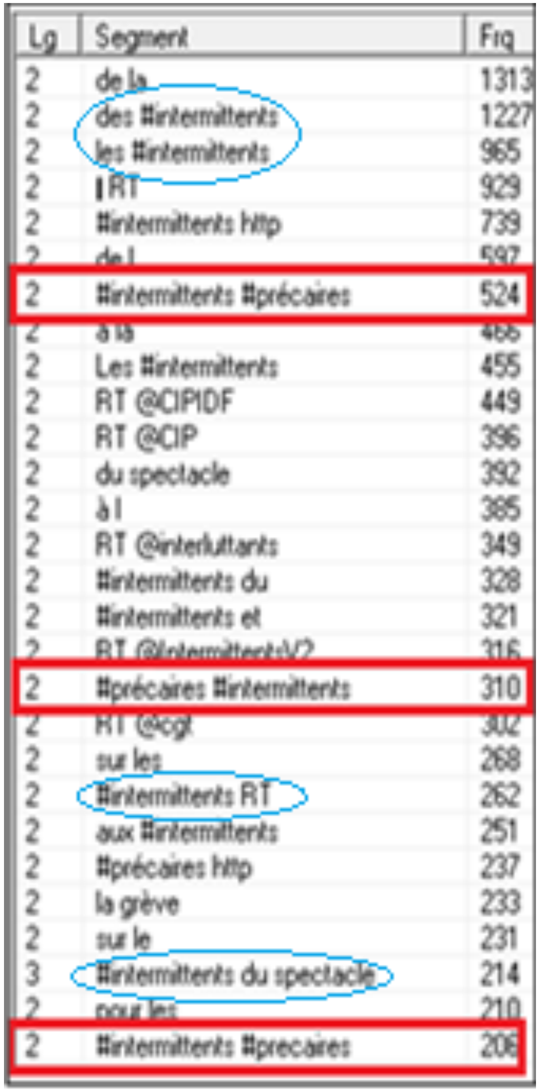
```
@mat _ gregoire : Quelle indemnisation pour les #intermittents du spectacle ?
application des droits rechargeables pour certains #intermittents ? cf . fin du doc
IU045uOU0V L ' interluttant : à tous , pour tous ! lire , imprimer , diffuser . . .
tous concernés pour un système juste pour TOUS les précaires ? http : / / t . co /
de concertations n'ont rien réglé , pour personne . Mais ont permis une chose : Le
lutte des #intermittents est une leçon pour l'ensemble du salariat» Entretien avec Mathieu
' espère qu ' une solution sera trouvée pour les #intermittents dans les prochaines 48
c ' est juste prendre les #intermittents pour des jambons ! @Culturebox @Mil _ 3000 oui
```

The hashtag

Used in the majority of tweets (of course #intermittent, but other)

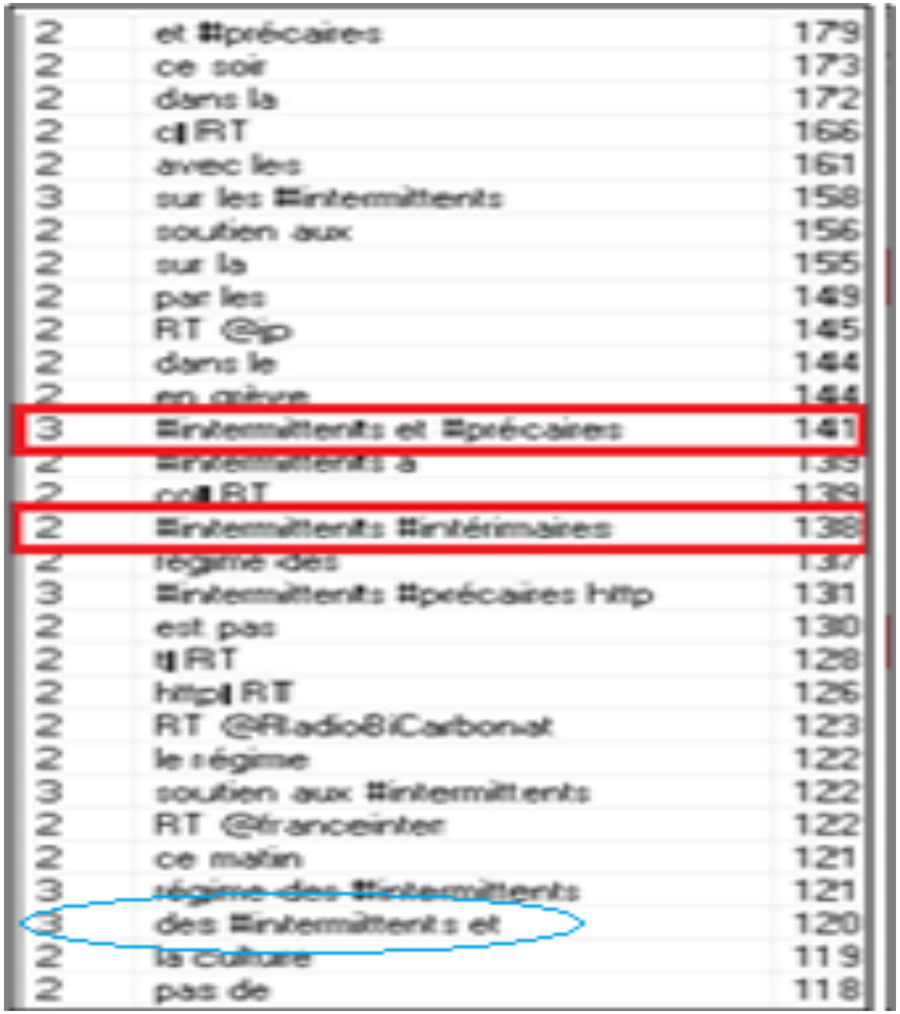
Lexemes

Cooccurrence of hastags



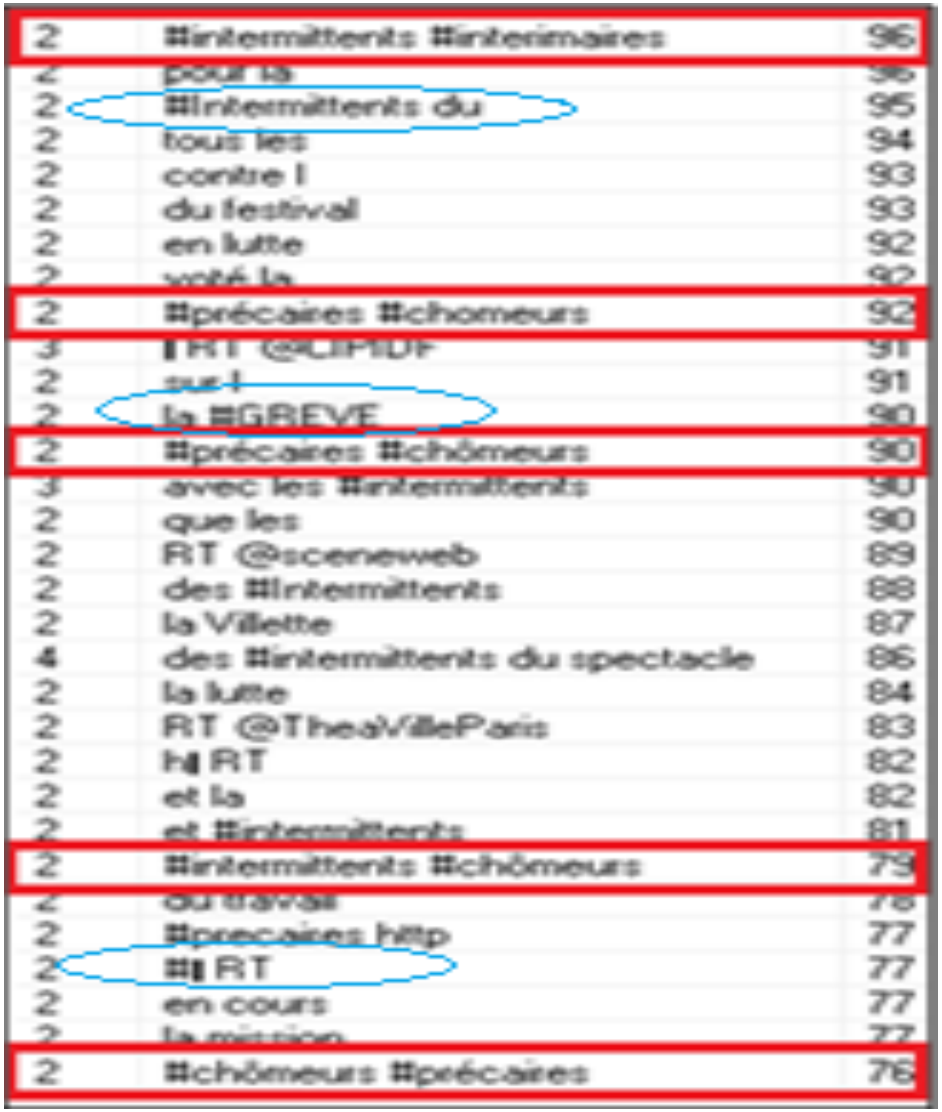
Lg	Segment	Frq
2	de la	1313
2	des #intermittents	1227
2	les #intermittents	965
2	| RT	929
2	#intermittents http	739
2	de l	597
2	#intermittents #précaires	524
2	à la	466
2	Les #intermittents	455
2	RT @CIPIDF	449
2	RT @CIP	396
2	du spectacle	392
2	à l	385
2	RT @interluttants	349
2	#intermittents du	328
2	#intermittents et	321
2	RT @IntermittentsV2	316
2	#précaires #intermittents	310
2	RT @cgt	302
2	sur les	268
2	#intermittents RT	262
2	aux #intermittents	251
2	#précaires http	237
2	la grève	233
2	sur le	231
3	#intermittents du spectacle	214
2	pour les	210
2	#intermittents #precaires	206

#intermittent(s) & #précaire(s)



2	et #précaires	179
2	ce soir	173
2	dans la	172
2	c▮ RT	166
2	avec les	161
3	sur les #intermittents	158
2	soutien aux	156
2	sur la	155
2	par les	149
2	RT @jp	145
2	dans le	144
2	en grève	144
3	#intermittents et #précaires	141
2	#intermittents à	139
2	coll▮ RT	139
2	#intermittents #intérimaires	138
2	régime des	137
3	#intermittents #précaires http	131
2	est pas	130
2	t▮ RT	128
2	http▮ RT	126
2	RT @RadioBiCarbonat	123
2	le régime	122
3	soutien aux #intermittents	122
2	RT @franceinter	122
2	ce matin	121
3	régime des #intermittents	121
3	des #intermittents et	120
2	la culture	119
2	pas de	118

#intermittent(s) & #intérimaire(s)



2	#intermittents #interimaires	96
2	pour la	96
2	#Intermittents du	95
2	tous les	94
2	contre l	93
2	du festival	93
2	en lutte	92
2	voté la	92
2	#précaires #chomeurs	92
3	▌RT @CIPIDF	91
2	sur l	91
2	la #GREVE	90
2	#précaires #chômeurs	90
3	avec les #intermittents	90
2	que les	90
2	RT @sceneweb	89
2	des #Intermittents	88
2	la Villette	87
4	des #intermittents du spectacle	86
2	la lutte	84
2	RT @TheaVilleParis	83
2	h▌ RT	82
2	et la	82
2	et #intermittents	81
2	#intermittents #chômeurs	79
2	du travail	78
2	#precaires http	77
2	#▌ RT	77
2	en cours	77
2	la mission	77
2	#chômeurs #précaires	76

#intermittent(s) & #chomeur(s)

The hashtag

Generally presented in two forms: either a lexeme used alone and returning more often adjectives (“*#précaires*” (precarious), “*#intérimaires*”(temporary)…) or names (“*#grève*” (strike), *#intermittents*” (arts workers) ...) or co-occurrences composed of two or more hashtags placed next to one another (“*#intermittents #précaires*”, “*#intermittents #chômeurs*”, “*#précaires #intermittents #interimaire*”...) and having the same nature as the lexemes listed above.

These different types of hashtag are placed either at the beginning, middle or at the end of a tweet. Cf Jackiewicz, 2014.

Conclusions & future works

These results demonstrate that unlike the written press which showed a plurality of views concerning the semantic representation of the word “*intermittent*” (see Longhi, 2006) which was seen whether as a status (*statut*), a profession (*métier*) or in the dynamics of these two semantic components. Here, the word “*intermittent*” is presented using three different senses "system" (*régime*), "status" (*statut*) and "fight” (*lutte*).

This indicates that Twitter focuses on the status side and declines it by introducing the French arts workers insurance system (one way of looking at the status) or the consequence of this status (fight).

A Textometrical analysis of this corpus has allowed us to see how twittos have reacted to the announcement of the new unemployment insurance system related to French arts workers.

There were also various retweets and thanks to this, the issue has become in a short time a "trending topic" on Twitter. This is due to the various markers such as #, URLs, the @...

The Reneirt method (HDC) taught us that discourse around this subject is divided into two different sets: precariousness of French arts workers & their various protest movements // the impartiality of the agreement, with links providing information.

Thank you