

Le corpus *Polititweets*

Enjeux de constitution d'un corpus de
tweets et propositions d'analyses

Julien Longhi, Université de Cergy-Pontoise

Plan de l'exposé

- Contexte des projets de recherche
- Méthodologie: pluridisciplinarité, sémioticité, complexité
- Premières analyses
- ... et difficultés
- Orientations et traitements récents
- Pistes et conclusions

Contexte des projets

...

Humanum_DJ et ICI 2.0

- Recherche pluridisciplinaire: initialement linguistique et informatique
- Analyse du lexique politique sur Twitter
- Mise en œuvre des méthodes et outils issus de l'analyse du discours: constitution de corpus, dimension sémiotique du langage, critères discursifs
- Usage des méthodologies et principes d'informatique: algorithmes, taux de rappel et de précision

Méthodologie

• • •

Pluridisciplinarité, sémioticité, complexité



Humanistica @HumanisticaDH · 7 mai

Création d'une collection "Humanités numériques" chez L'Harmattan
fr.groups.yahoo.com/neo/groups/par... via @jlonghi1



8



5



Humanistica a Retweeté



Solène Roulet-Pillon @SoleneRoulet · 5 mai

Atelier : création d'une communauté en arts et humanités numériques |
THATCamp Paris tcp.hypotheses.org/836



2



2



Humanistica @HumanisticaDH · 2 mai

5. Vos droits

Vous conservez vos droits sur tous les Contenus que vous soumettez, postez ou publiez sur ou par l'intermédiaire des Services. En soumettant, publiant ou affichant des Contenus sur ou par le biais des Services, vous nous accordez une licence mondiale, non exclusive, gratuite, incluant le droit d'accorder une sous-licence, d'utiliser, de copier, de reproduire, de traiter, d'adapter, de modifier, de publier, de transmettre, d'afficher et de distribuer ces Contenus sur tout support par toute méthode de distribution connue ou amenée à exister.



Attention : cette licence signifie que vous nous autorisez à mettre vos Tweets sur les Services Twitter à la disposition du reste du monde et que vous permettez aux autres d'en faire de même.

Vous consentez à ce que cette licence comprenne le droit pour Twitter de fournir, de promouvoir et d'améliorer les Services et de mettre les Contenus publiés ou transmis au travers des Services à disposition d'autres sociétés, organisations ou individus en partenariat avec Twitter pour l'agrégation, la diffusion, la distribution ou la publication de ces Contenus sur d'autres supports, médias et services, dans la limite de nos conditions pour l'utilisation de ces Contenus.



Attention : Twitter applique un ensemble évolutif de règles sur la manière dont les partenaires de l'écosystème peuvent interagir avec vos Contenus sur les Services Twitter. Ces règles ont été conçues pour mettre en place un écosystème ouvert, tenant compte de vos droits. Mais ce qui vous appartient vous appartient, vous restez propriétaire de vos Contenus (et vos photos font partie de ces Contenus).

Ces usages supplémentaires par Twitter, ou d'autres sociétés, organisations ou individus en partenariat avec Twitter, peuvent être faits sans compensation à votre égard en ce qui concerne les Contenus que vous soumettez, publiez, transmettez ou rendez disponibles au travers des Services.

BDD

tweet_text	screen_name	favoritecount	retweetcount
Superbe photo et geste de soutien des spectateurs ...	CIP_LR	19	71
Intervention des représentants des #intermittents ...	IntermittentsV2	15	34
À propos des intermittents, dans l'émission de dim...	JLMelenchon	15	39
Le régime des #intermittents : tout comprendre en ...	franceinter	15	25
Jour de carence des fonctionnaires, #intermittents...	exbabacool	13	33
Bravo cher @PatSebastien Voilà quand un homme de c...	maximedereymez	13	12
#intermittents Les Conseillers régionaux PSR d'Île...	ELUSPSCRIDF	11	37
Lettre à @FHollande au sujet des #intermittents de...	IntermittentsV2	11	40
Au nom de quoi le manutentionnaire ou la caissière...	abrezet	10	31
Opéra national de #Lyon ce matin #intermittents #C...	jcgaleazzi	10	19
Comment les media ont-ils traité des #intermittent...	FredMetzo	10	14
Les #intermittentsduspectacle en ce moment!! @Toul...	GGBIANCHINI	9	14
Un troll pour expliquer ce besoin systématique de ...	jeunefrancaise	9	39
On peut penser ce qu'on veut de @JLMelenchon mais ...	IntermittentsV2	9	5
Avec #EELV et le #FDG je vote une motion du @Regio...	pbeauvillard	9	23
Vu dans la manif #intermittents tellement juste : (...	julienbayou	9	21
#Now Occupation du théâtre de la Ville à Paris par...	OlivierCyran	9	31
Les #intermittents sont comme les abeilles, leur d...	JulioFlo31	8	22
#intermittents "Votre statut est une chance, prot...	franceinter	8	19
Reportage : à Avignon, la question des #intermitte...	Telerama	8	24
À lire : l'interview poignante d'Olivier Py dans @...	CecileDuflot	8	8
Excellent article de @lesinRocks - #intermittents,...	IntermittentsV2	8	18
Mon Rapport de mission « Une nouvelle donne pour ...	jp_gille	8	31
Après une représentation, une personne de la #Comé...	pchenebenoit	8	22
#intermittents : un accord pour qui ? Demain à 7h4...	Lesmatinsfcult	8	20
"Jouer pour soutenir" @LoicCorbery et la #troupe d...	FaustineRoche	8	6
Avignon -Place Pie #intermittents #FA2014 http://t...	chauvetdavid	8	30

Constitution du corpus

- 1) Partir de 7 personnalités de **6 groupes politiques**
- 2) Récupérer toutes les listes créées par des utilisateurs où ils étaient cités : cela a conduit à **7087 listes** ;
- 3) Sélectionner parmi ces **listes** celles qui avaient **au moins 6 twittos** (utilisateurs de Twitter) et qui contenaient la chaîne de caractère ***politic*** dans le nom ou descriptif de la liste : 120 listes ;
- 4) Sur ces 120 listes récupérer les **2934 twittos** ;
- 5) **Travail par seuil**, en ne retenant que les comptes présents dans plus de 12 listes : nous arrivons ainsi à **205 twittos politiques**.
- Sur ces 205 comptes nous avons récupéré les **200 derniers tweets de chacun** au 27 mars 2014, soit 34273 tweets.

Comment télécharger un corpus CoMeRe ?

Prenons l'exemple de ce corpus de Tweets (pour la liste complète des corpus CoMeRe de différents genres, voir <http://hdl.handle.net/11403/comere>) :

- Longhi, J., Marinica, C., Borzic, B. & Alkhouli, A. (2014). Polititweets, corpus de tweets provenant de comptes politiques influents. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy.
<http://hdl.handle.net/11403/comere/cmr-polititweets>

En cliquant sur ce lien <http://hdl.handle.net/11403/comere/cmr-polititweets>, vous accédez à la fiche descriptive général du corpus (fiche OLAC de métadonnées).



Corpus



This page: <http://hdl.handle.net/11403/comere/cmr-polititweets>

Back to Repository main page: <http://hdl.handle.net/11403/comere>

Corpus CoMeRe cmr-polititweets-tei-v1 : corpus Polititweet, tweets provenant de comptes politiques influents 1

How to cite this resource

Longhi, J., Marinica, C., Borzic, B., Alkhoul, A.(2014). Polititweets : corpus de tweets provenant de comptes politiques influents 1. In Chanier T. (ed) Banque de corpus CoMeRe. Ortolang.fr : Nancy. [<http://hdl.handle.net/11403>]

Description

The corpus Polititweets gathers tweets of 7 personalities from 6 French different political groups : Mélenchon, Bayrou, Copé, Fillon, Le Pen, Ayrault, Cohn-Bendit. Extracted from 205 personalities, by a method that selected the messages of 205 twittos send in 2013-14, what makes for the total corpus 34273 messages (tweets) that contain 502 085 tokens / words. The corpus is related to the campaign of municipal elections of March 2014. The first version of this corpus has been build up by the project "Numerical humanities and data journalism". The initial corpus initial has been converted to the TEI format within the framework of the project CoMeRe (Communication médiée par les réseaux, Network mediated communication) <http://hdl.handle.net/11403/comere>

The complete corpus consists in 7 TEI folders. In the body of each folder, the messages coming from about 30 user accounts / twittos are gathered. The first serie of messages come from 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. The project received the support of ORTOLANG (the French equivalent of DARIAH) and of the national consortium Written-Corpus ('Corpus-écrits', <http://corpusecrits.corpus-ir.fr>), subsection of H

Keywords: Tweet, Computer Mediated Communication, CMC,

- Created on: 2014-06-04
- Language: fra
- Coverage: 205 user accounts / twittos ; 34273 posts / tweets ; 567851 tokens
- Time of data collection: name=polititweets ; start=2009-03-04 ; end=2014-03-27
- ConformTo: TEI (Text Encoding Initiative)The TEI structure used is an extension of TEI for CMC genres. This extension is developed by a European project for which the participants are : Michael Beißwenger¹ (DE), Thier 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



Parti de Gauche (PG) @LePG · 25 mars 2014

A #Bruxelles, @JLMelenchon conclue la rencontre-débat sur le #GMT.
Nous live-tweetons - #UE #USA #Europe -



← ↻ 21 ★ 3 ...

[Voir la photo](#)


```

550 |
551 | <tweet_id="448491380948885504">
552 | <author>
553 |   <user_id>
554 |     | 80820758
555 |   </user_id>
556 |   <name>
557 |     | Jean-Luc Mélenchon
558 |   </name>
559 |   <screen_name>
560 |     | JLMelenchon
561 |   </screen_name>
562 | </author>
563 | <creation_date>
564 |   | 2014-03-25 17:07:32.0
565 | </creation_date>
566 | <tweet_text>
567 |   | RT @LePG: A: #Bruxelles: @JLMelenchon conclue la rencontre-débat sur le #GMT. Nous live-tweetons: #UE #USA #Europe: http://t.co/hNGmnAagI
568 | </tweet_text>
569 | <entities_hashtags>
570 |   | Bruxelles,GMT,UE,USA,Europe
571 | </entities_hashtags>
572 | <entities_urls>
573 |   |
574 | </entities_urls>
575 | <entities_user_mentions>
576 |   | 35022597,80820758
577 | </entities_user_mentions>
578 | <geo_lat>
579 |   | 0.0
580 | </geo_lat>
581 | <geo_long>
582 |   | 0.0
583 | </geo_long>
584 | <source>
585 |   | <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
586 | </source>

```

TEI: Text Encoding Initiative

The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form. Its chief deliverable is a set of Guidelines which specify encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation. In addition to the Guidelines themselves, the Consortium provides a variety of [resources](#) and [training events](#) for learning TEI, information on [projects using the TEI](#), a [bibliography of TEI-related publications](#), and [software](#) developed for or adapted to the TEI.

The TEI Consortium is a nonprofit membership organization composed of academic institutions, research projects, and individual scholars from around the world. Members contribute financially to the Consortium and elect representatives to its Council and Board of Directors.

halshs-01176061, version 1

Structuring a CMC corpus of political tweets in TEI: corpus features, ethics and workflow

Julien Longhi ¹, Ciara R. Wigham ² [Détails](#)

- 1** EA 1392 CRTF
CRTF - Centre de recherche textes et francophonies
- 2** ICAR - Interactions, Corpus, Apprentissages, Représentations

Abstract : The CoMeRe project (CoMeRe, 2014) 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. The project gathered mono- and multimodal, synchronous and asynchronous communication data from both Internet and telecommunication networks (text chat, tweets, SMSs, forums, blogs). A variety of interactions was sought: public or private interactions as well as interactions from informal, learning and professional situations. Whereas some CMC data types were collected within the CoMeRe project, others had previously been collected and structured within different project partners' local research teams. This meant that the project had to overcome disparities in corpus compilation choices. For this reason, the CoMeRe project structured the corpora in a uniform way using the Text Encoding Initiative format (TEI, Burnard & Bauman, 2013) and decided to describe 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



Parti de Gauche (PG) @LePG · 25 mars 2014

A #Bruxelles, @JLMelenchon conclue la rencontre-débat sur le #GMT.
Nous live-tweetons - #UE #USA #Europe -



21



3



[Voir la photo](#)


```

<post xml:id="cmr-politweets-a448491380948885504" who="#cmr-politweets-p80820758"
  when="2014-03-25T17:07:32" xml:lang="fra">
  <p><distinct type="twitter-retweet"><ident>RT</ident>
    <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
      ref="https://twitter.com/LePG 35022597"
    >LePG</addressee></addressingTerm>:</distinct> A <distinct type="twitter-hashtag"
      ><ident>#</ident><rs ref="https://twitter.com/search?q=%23Bruxelles&src=hash"
      >Bruxelles</rs></distinct>,
      <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
        ref="#cmr-politweets-p80820758">JLMelenchon</addressee></addressingTerm> conclue
la rencontre-débat sur le <distinct type="twitter-hashtag"><ident>#</ident><rs
  ref="https://twitter.com/search?q=%23GMT&src=hash">GMT</rs></distinct>. Nous
live-tweetons - <distinct type="twitter-hashtag"><ident>#</ident><rs
  ref="https://twitter.com/search?q=%23UE&src=hash">UE</rs></distinct>
<distinct type="twitter-hashtag"><ident>#</ident><rs
  ref="https://twitter.com/search?q=%23USA&src=hash">USA</rs></distinct>
<distinct type="twitter-hashtag"><ident>#</ident><rs
  ref="https://twitter.com/search?q=%23Europe&src=hash">Europe</rs></distinct> -
  <ref target="http://t.co/hNGmnAagIM">http://t.co/hNGmnAagIM</ref></p>
<trailer>
  <fs>
    <f name="medium">
      <string>Twitter for iPhone</string>
    </f>
    <f name="retweetcount">
      <numeric value="22"/>
    </f>
    <f name="isRetweet">
      <binary value="true"/>
    </f>
    <f name="retweetedstatus_id">
      <numeric value="448490939338588160"/>
    </f>
  </fs>
</trailer>
</post>

```

```

<post xml:id="cmr-politweets-a448491380948885504" who="#cmr-politweets-p80820758"
  when="2014-03-25T17:07:32" xml:lang="fra">
  <p><distinct type="twitter-retweet"><ident>RT</ident>
    <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
      ref="https://twitter.com/LePG 35022597"
    >LePG</addressee></addressingTerm>:</distinct> A <distinct type="twitter-hashtag"
      ><ident>#</ident><rs ref="https://twitter.com/search?q=%23Bruxelles&src=hash"
    >Bruxelles</rs></distinct>,
    <addressingTerm><addressMarker>@</addressMarker><addressee type="twitter-account"
      ref="#cmr-politweets-p80820758">JLMelenchon</addressee></addressingTerm> conclue
    la rencontre-débat sur le <distinct type="twitter-hashtag"><ident>#</ident><rs
      ref="https://twitter.com/search?q=%23GMT&src=hash">GMT</rs></distinct>. Nous
    live-tweetons - <distinct type="twitter-hashtag"><ident>#</ident><rs
      ref="https://twitter.com/search?q=%23UE&src=hash">UE</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
      ref="https://twitter.com/search?q=%23USA&src=hash">USA</rs></distinct>
    <distinct type="twitter-hashtag"><ident>#</ident><rs
      ref="https://twitter.com/search?q=%23Europe&src=hash">Europe</rs></distinct> -
    <ref target="http://t.co/hNGmnAagIM">http://t.co/hNGmnAagIM</ref></p>
  <trailer>
    <fs>
      <f name="medium">
        <string>Twitter for iPhone</string>
      </f>
      <f name="retweetcount">
        <numeric value="22"/>
      </f>
      <f name="isRetweet">
        <binary value="true"/>
      </f>
      <f name="retweetedstatus_id">
        <numeric value="448490939338588160"/>
      </f>
    </fs>
  </trailer>
</post>

```

The @

The #

Type of material

Is retweeted/ number of
retweets

Extracts of Interactions

head:Tweets de Jean-Luc Mélenchon

- **POST:** xml:id: p_ts-a449071343175471104 | who: s-p80820758 | when: 2014-03-27T07:32:06 | xml:lang: fra
p: RT @LePG: A 07h50, @JLMelenchon est l'invité des #4vérités sur #France2. Nous live-tweeterons. #Chômage #Municipales2014
trailer:
 - medium: web
 - retweetcount: 8
 - isRetweet: true
 - retweetedstatus_id: 449071255489372160
- **POST:** xml:id: p_ts-a448882215209144320 | who: s-p80820758 | when: 2014-03-26T19:00:34 | xml:lang: fra
p: Ce jeudi à 07h50, je suis l'invité des #4vérités sur @France2tv. Live-tweet sur @LePG. #Chômage #Municipales2014
trailer:
 - medium: web
 - favoritecount: 4
 - retweetcount: 35

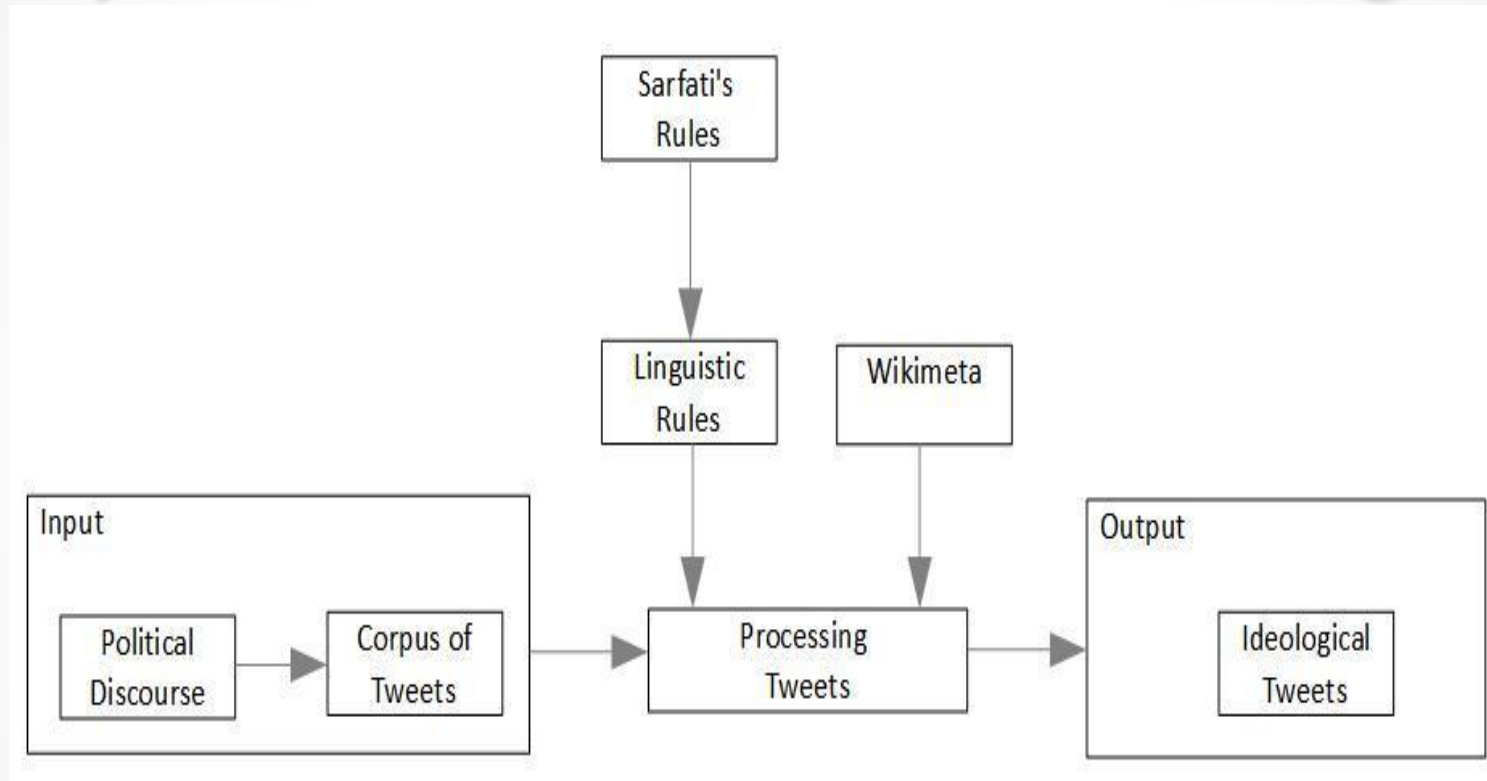
head:Tweets de Chantal Jouanno

- **POST:** xml:id: p_ts-a448872049034141697 | who: s-p102397481 | when: 2014-03-26T18:20:10 | xml:lang: fra
p: Si le journaliste sur @RTLFrance pouvait organiser la prise de parole, on comprendrait mieux le débat
trailer:
 - medium: Twitter for iPhone
 - favoritecount: 7
 - retweetcount: 26
- **POST:** xml:id: p_ts-a448771540658978816 | who: s-p102397481 | when: 2014-03-26T11:40:47 | xml:lang: fra
p: Réunion de travail @Les_Europeens sur notre projet. Une vraie chance pour notre pays. <http://t.co/W39cSzcO2L>
trailer:
 - medium: Twitter for iPhone
 - favoritecount: 7
 - retweetcount: 22

Premières analyses ... et difficultés

...

Projet Recherche d'idéologies



*Processus de traitement du corpus à partir des règles établies (Djemili, Longhi et al. 2014:
What does Twitter have to say about ideology?)*

Exemples



Guillaume Peltier

@G_Peltier



Suivre

La force de la volonté rend possible l'action et permet de soulever les montagnes. Les solutions, on les connaît, il manque le courage...



RETWEETS

28

FAVORIS

4



21:12 - 28 nov. 2013



Faouzi Lamdaoui @Faouzi_Lamdaoui · 2 févr. 2014

La politique de solidarité avec les familles s'inscrit pleinement dans la fraternité républicaine qui doit unir tous les citoyens.

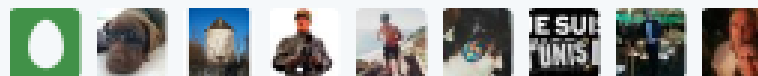


RETWEETS

22

FAVORIS

9



18



Faouzi Lamdaoui ✓

@Faouzi_Lamdaoui



Suivre

La diversité est une réalité structurelle de la société française depuis ses origines. Parler d'[#integration](#) est un non-sens sociologique.

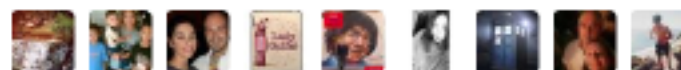


RETWEETS

466

FAVORIS

135



12:27 - 8 févr. 2014

Implémentation de l'approche

■ Implémentation des règles

R 1 : pas de deixis spatio-temporelle

NDAY, NMONTH, NTIME tags
17 spatio-temporal markers :
"maintenant", "demain"...

R 2 : pas de sujets d'interlocution

PRO tag
"je", "tu", "nous", "vous", "moi" ...

R 3 : pas de noms propre

XNAM, ABR tags

R 4 : pas de modalisations, pas de
"!" / "?"

Lemmatization API (NLP group of Sheffield University)
"sembler", "paraître", "pouvoir", "devoir"
"!" and "?"

R 5 : pas de connecteurs argumentatifs

KON tag
Usual argumentative connectors:
"mais", "or", "puisque",...

R 6 : présent dominant

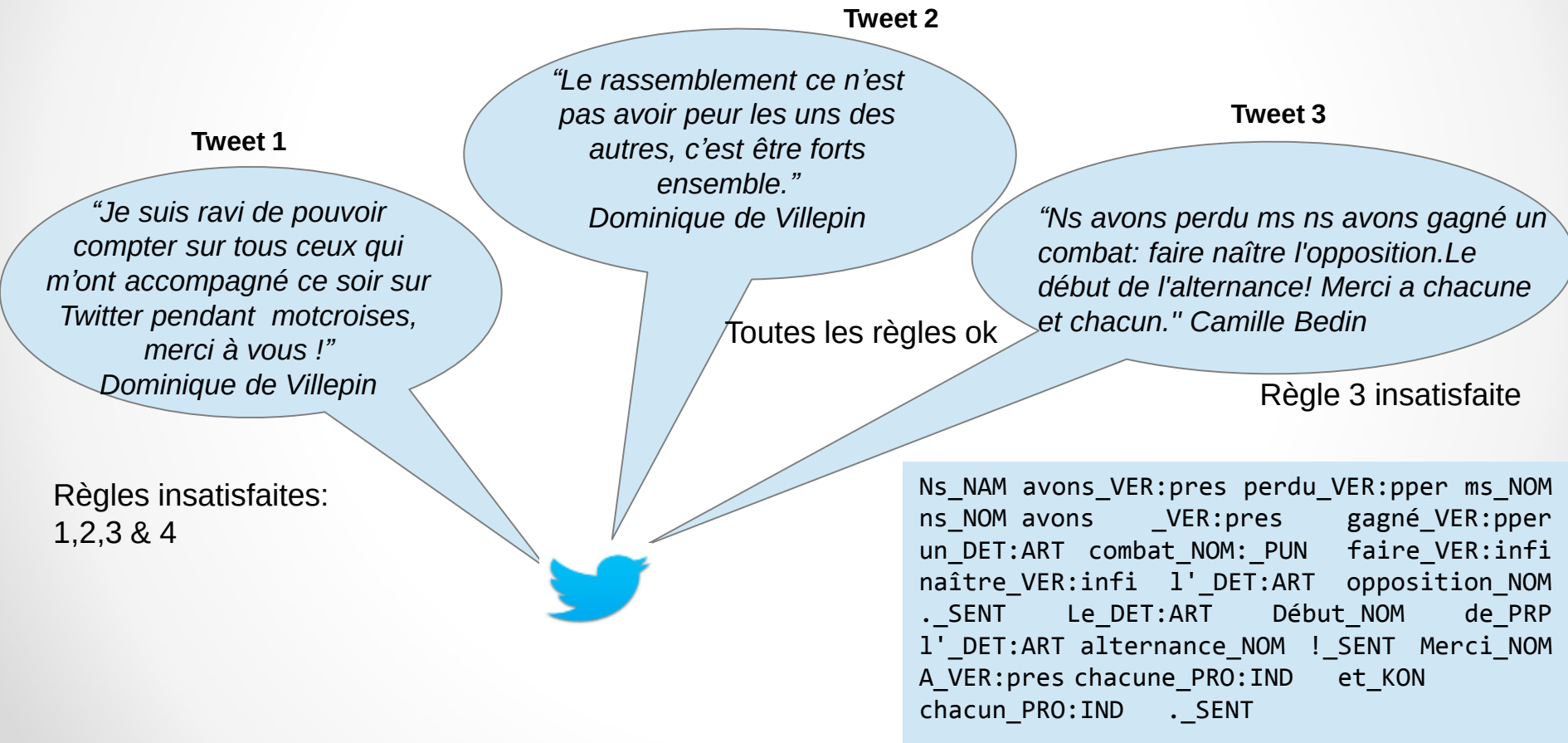
VER tag
cpt1 = past, cpt2= present, cpt3= future

R 7 : pas de marques du discours

(Giguet et Lucas, 2001) approach:
PERS and VER:pper tags, "...", "selon"
+ @: reply

Implémentation de l'approche (2)

■ Application des règles



Résultats et discussions(1)

Analyses des “faux positifs”

Nombre de tweets idéologiques détectés	% d'idéologiques après validation (Precision)	% de non-idéologiques après validation	% de “partiellement” idéologiques après validation
321 sur les 20400	66.66	23.36	9.96

■ Non-idéologiques:

@ : *@askolovitchC il faut conduire avec moderation... // @lasserrelo ca coule de source...*

: *#retraites : visiblement on s'oriente vers du grand n'importe quoi ...*

■ “Partiellement” idéologiques:

#Confsociale : l'uniformisation et la simplification des systèmes de prévention sociale et de retraite s'impose dès à présent.

@DominiqueReynie bravo pour ce travail. l'innovation est forcément une contestation de l'existant

■ Tweets avec @ ou # mais validés comme idéologiques:

Le progrès social n'est pas l'adversaire de la performance économique #loiESS

Résultats et discussions(2)

Analyse des faux négatifs

■ Non-idéologiques par le système: 20079 tweets

Sélection par seuils:

- 4% des tweets qui ne satisfaisaient pas 1 règle (117 tweets)
- 2% des autres tweets (329 tweets).

Total des tweets analysés	% des tweets correctement classés comme idéologiques	% de faux-négatifs
446	96.64	3.36

■ Pas d'erreurs de classification si 3 règles ou plus n'étaient pas satisfaites: toujours correctement identifié comme non-idéologique

Résultats et discussions(3)

Analyse des faux négatifs

■ Pourquoi sont-ils mal classés par le système?

(1) Erreur d'annotation de Wikimeta

(2) Règle 2: parfois les sujets de l'interlocution (nous) sont utilisés pour des référents généraux

(3) Règle 6: le futur peut avoir une valeur prospective idéologique

La République sera à tous les Français

Résultats et discussions(4)

■ Des structures/patrons ont été identifiés

- “Il faut que”

Ce qu’il faut c’est établir des priorités, choisir des filières d’excellence, créer des emplois dans des secteurs porteurs.

- “Il y a”

Il y a un problème de méthode pour régler les problèmes que rencontrent nos banlieues ; il faut développer des conseils de quartier élus.

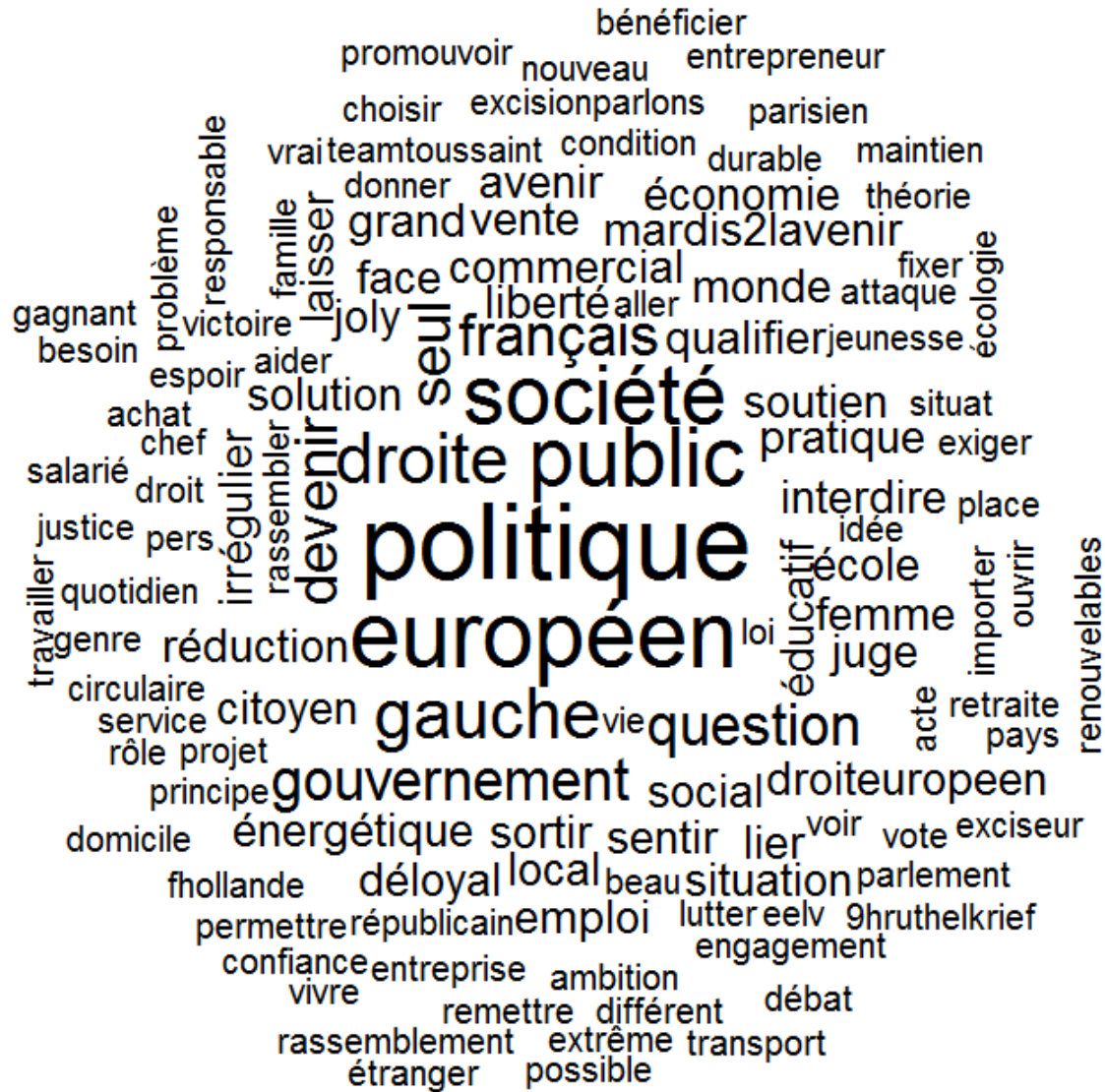
- Focalisation (X, c’est x // ce qui est...c’est)

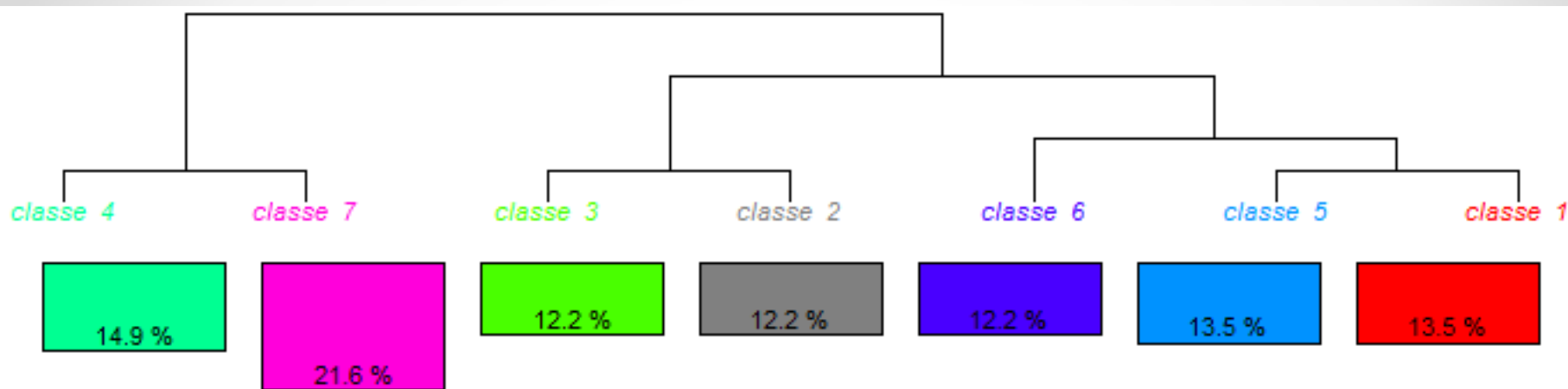
Promouvoir un dialogue entre les équipes éducatives, les élèves et les parents, c’est ouvrir la voie à une école bienveillante.

Rebond: des « golden » corpus

- Tweets idéologiques:
 - 172 tweets
 - Traitement dans Iramuteq
- Tweets efficaces
 - 600 tweets
 - Traitement dans Iramuteq
- Tweets négatifs
 - 3190 tweets, extraits par Unitex
 - Ebauches de traitements

Idéologiques





joly
monde
vivre
possible
permettre
excisionparlon:
espoir
eelv
droit
acte
solution
responsable
problème
différent
avenir
français
grand
femme

société
soutien
vie
teamtoussaint
rôle
retraite
quotidien
nouveau
maintien
lutter
exiger
entrepreneur
domicile
besoin
aller
aider
achat
9hruthelkrief
gouvernement
français
femme
idée
salarié
loi
justice
confiance
solution

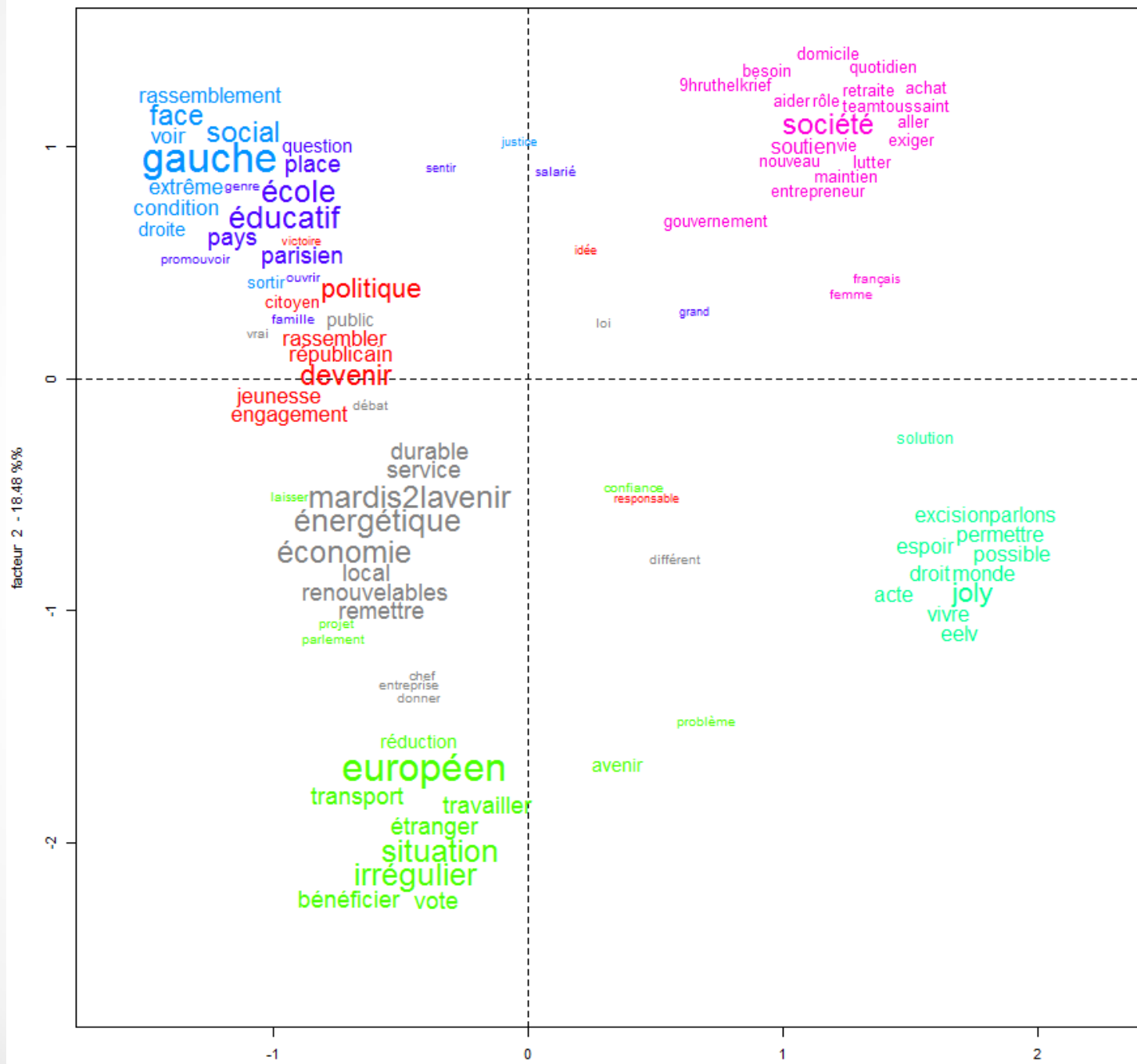
européen
situation
irrégulier
étranger
vote
travailler
transport
bénéficier
réduction
avenir
projet
parlement
laisser
problème
entreprise
donner
chef
confiance

énergétique
économie
mardis2laver
local
service
renouvelables
remettre
durable
public
vrai
loi
entreprise
débat
donner
différent
chef
sortir
réduction
gouvernement

éducatif
école
place
pays
parisien
public
question
droite
ouvrir
famille
vrai
salarié
promouvoir
genre
débat
grand
sentir

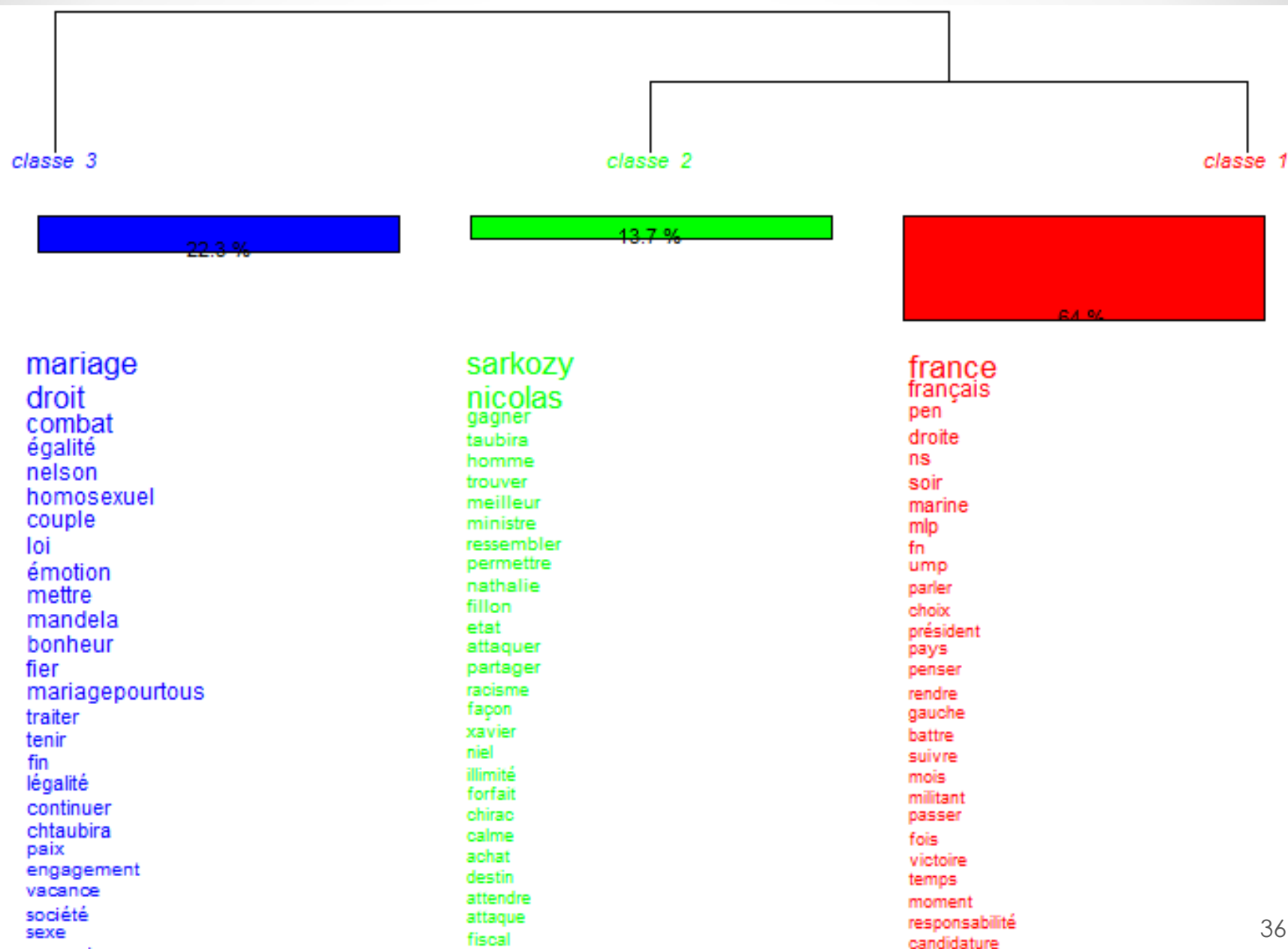
gauche
social
face
voir
rassemblement
extrême
condition
droite
sortir
victoire
laisser
promouvoir
justice
genre
citoyen
question
sentir
politique

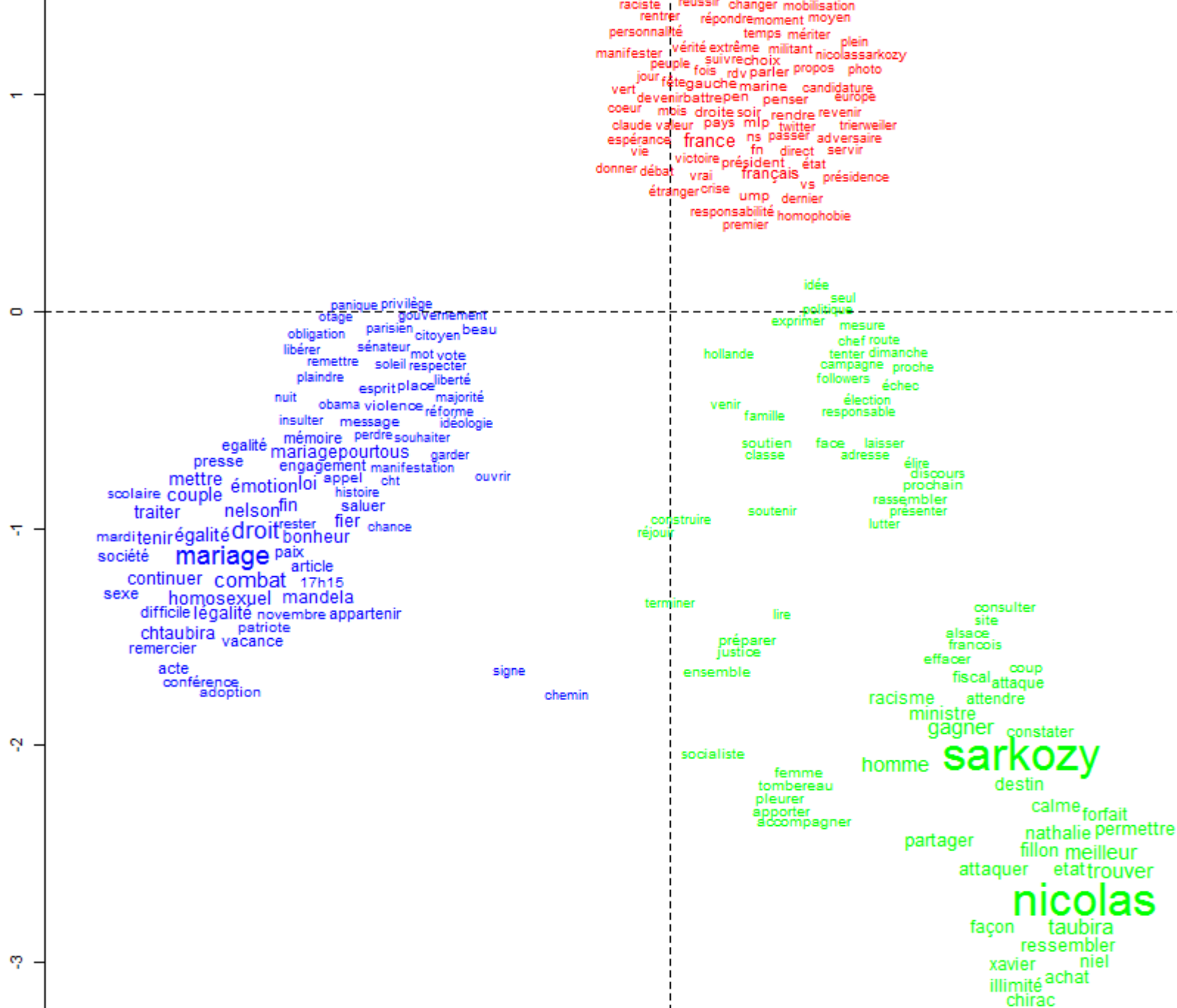
devenir
politique
républicain
rassembler
jeunesse
engagement
citoyen
victoire
responsable
projet
parlement
ouvrir
idée
famille

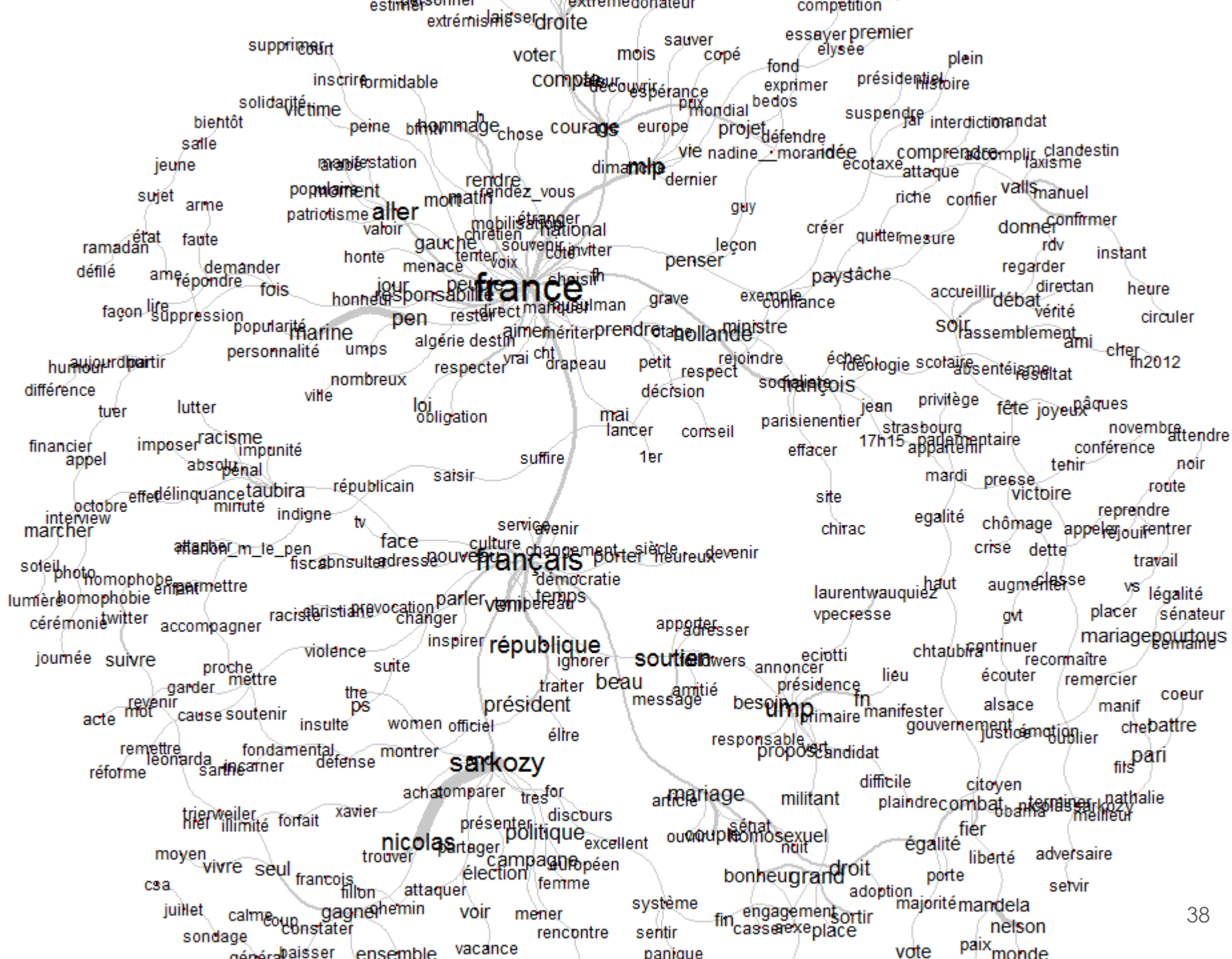


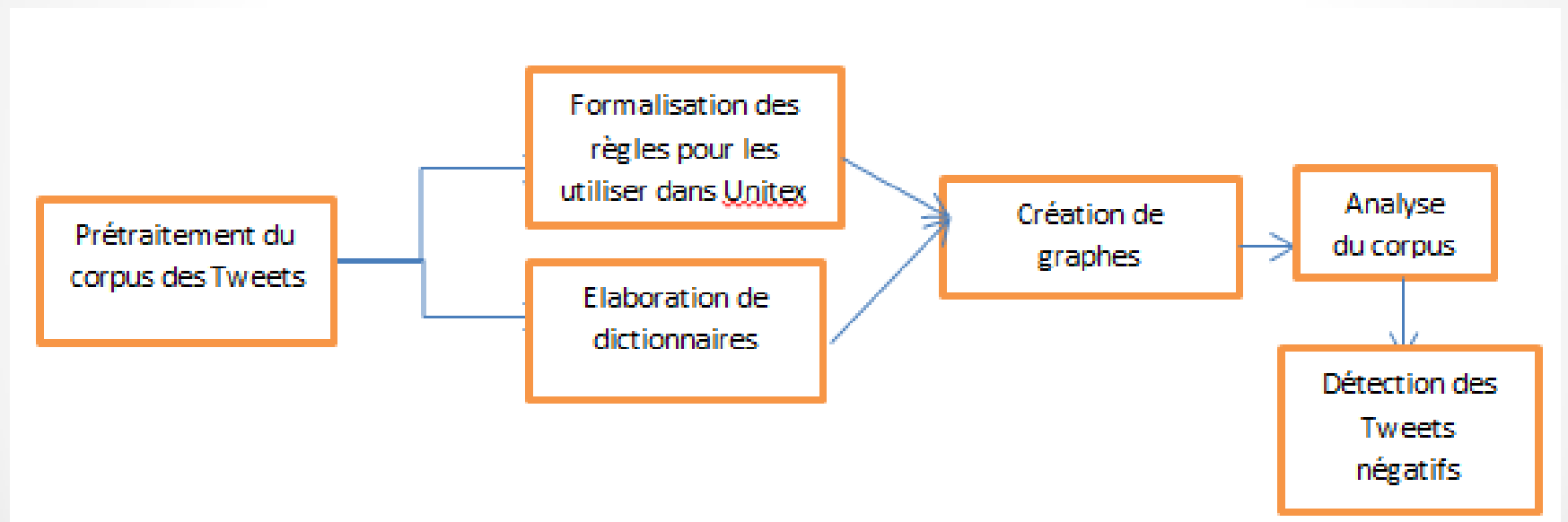
Tweets efficaces

tweet_text	screen_name	favoritecount ▼	retweetcount
Toute ma gratitude va à l'extraordinaire personnel ...	valtrier	2286	6337
Merci du fond du cœur à tous ceux qui ont envoyé d...	valtrier	1830	3429
"Soyons dignes, soyons patriotes, soyons Français ...	NicolasSarkozy	1632	5711
Avant d'entrer à la présidence de la République, je...	fhollande	1401	6806
Que chacun d'entre vous sache combien je suis reco...	NicolasSarkozy	1350	2617
"Vous pourrez compter sur moi à chaque fois qu'il y...	NicolasSarkozy	614	1832
"Jamais je ne pourrai vous rendre tout ce que vous...	NicolasSarkozy	499	2499
@nicodomenach je pourrais aussi entrer en arrière ...	ramayade	464	2979
Joie d'accueillir le père Georges, rayonnant de gén...	fhollande	446	1129
#FF @Elysee pour suivre toutes les informations su...	fhollande	427	993











Exemples de règles

Règle1 : Cette règle regroupe les formes de négation verbale qui se construisent à partir de *ne*, suivi d'un verbe suivi des différents opérateurs de négation syntaxiques : *comme, pas, plus, rien, personne, jamais, aucun/aucune, nul/nulle*.

Ainsi, nous créons un graphe comportant la négation verbale classique englobant deux éléments principaux :

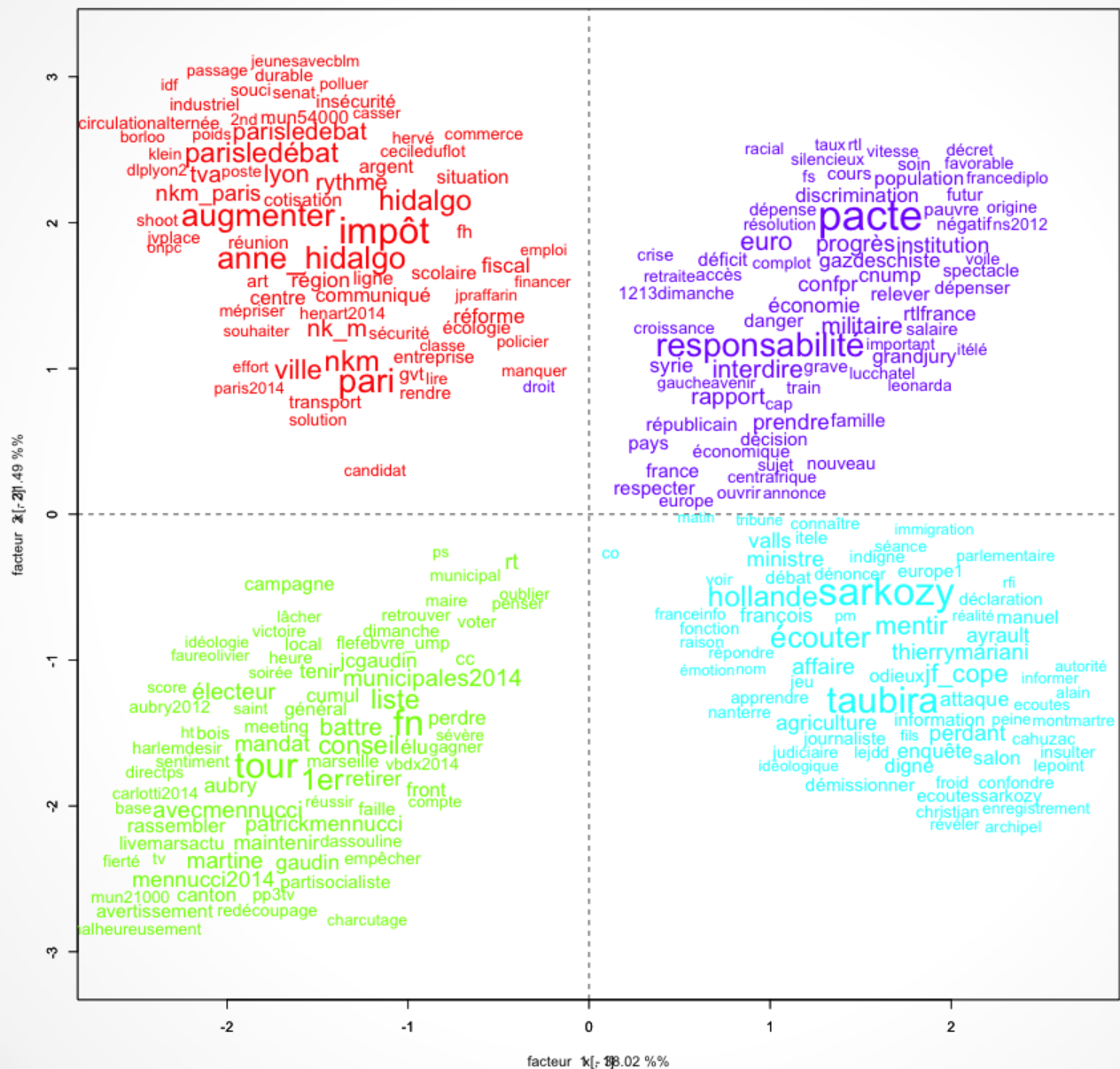
la négation verbale d'un verbe conjugué
<N+PRO>+<ne+n'>+<V>+<pas+plus+rien++personne+jamais+aucun/aucune+nul/nulle>, et
la négation qui entraîne un infinitif après les morphèmes :

<ne+n'>+<pas+jamais+rien+plus>+<V : W>.

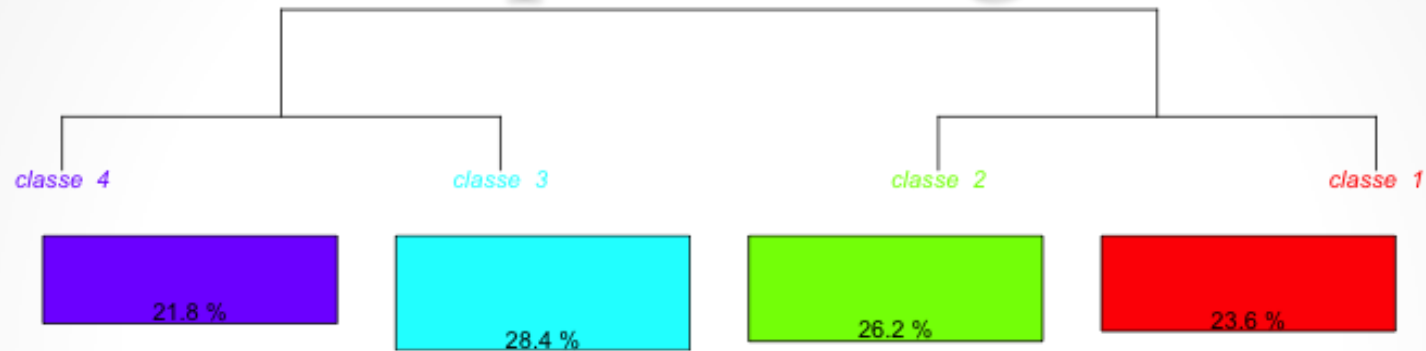
Règle 4 : Pour cette règle, nous relevons manuellement tous les prédicats à connotations négatives, et nous les formalisons dans le dictionnaire DELAF pour mots simples. Afin de permettre à notre outil de les détecter, nous rajoutons le code sémantique <NEG> au code <N> ou <V>, comme dans l'exemple :

Refuse, refuser.<V>+<NEG>

Règle	% des tweets satisfaisant la règle	Nombre de tweets satisfaisant la règle
R1	51%	1630
R2	1%	31
R3	2%	115
R4	40%	1290
R5	27%	864
R6	17%	558
R7	0%	0
R8	1%	32
R9	0,5%	19



Liens thématiques/négations?



pacte
responsabilité
euro
interdire
militaire
progrès
rapport
prendre
institution
confpr
économie
gazdeschiste
cnump
syrie
discrimination
grandjury
rtlfrance
respecter
relever
population
pays
france
danger
déficit
famille
républicain
décision
pauvre
négatif
futur
dAnenar

sarkozy
taubira
écouter
hollande
mentir
jf_cope
perdant
affaire
thierrymariani
ministre
agriculture
valls
ayrault
enquête
françois
attaque
buisson
digne
salon
manuel
odieux
démissionner
information
affirmer
débat
journaliste
europe1
ecoutessarkozy
ilele
discrimination

fn
tour
1er
liste
conseil
battre
municipales2014
électeur
avecmennucci
martine
mandat
rt
mennucci2014
patrickmennucci
retirer
aubry
gaudin
maintenir
tenir
élu
jcgauvin
perdre
cumul
front
rassembler
canton
campagne
cc
discrimination

impôt
pari
augmenter
anne_hidalgo
nkm
hidalgo
parisledébat
ville
lyon
parisledébat
nk_m
rythme
tva
nkm_paris
région
fiscal
réforme
communiqué
centre
situation
argent
ligne
scolaire
gvt
fn
mun54000
insécurité
entisation

Conclusion et perspectives

...

Méthodologie

- Prendre en compte la complexité de l'objet et les niveaux d'analyse possibles
- Complémentarité pluridisciplinaire: corpus, analyses, critères de validité
- Analyse « pragmatique », aller du corpus aux résultats, des résultats au corpus etc

Analyses

- Articulation des aspects traditionnellement qualifiés de qualitatifs et quantitatifs
- Intérêt des outils lexicométriques/textométriques: faire émerger des propriétés des corpus
- Mais intérêt d'analyses plus « volumineuses » même si moins linguistiques pour élaborer des hypothèses

Perspectives

- Question des outils et de leur application au format
- Liage des niveaux de l'expression et du contenu (*Théorie des objets discursifs*, J. Longhi 2015)
- Reproductibilité de l'analyse et phénomènes (cf. corpus #intermittent)