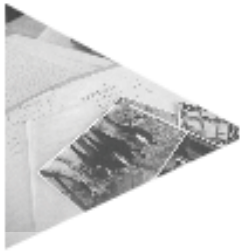
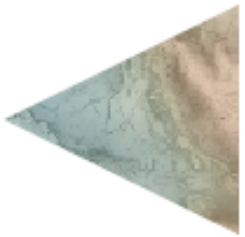
Twitter comme “ corpus ” en sciences du langage : questions méthodologiques et pistes de recherche

Laurent Gautier



Twitter comme « corpus » en sciences du langage : questions méthodologiques et pistes de recherche

Laurent Gautier, Centre Interlangues Texte Image Langage (UBFC, EA 4182) & MSH Dijon (USR uB – CNRS 3516)



Structure de la communication

1. Contexte
2. Problématique et objectifs
3. Quelles données ? Le défi de la constitution du corpus
4. Quels cadres méthodologiques ?
5. Discussion
6. Perspectives

1. Contexte

La linguistique ? de corpus…

- Résultat du **changement de paradigme** de la recherche en sciences du langage => linguistique de la parole vs. de la langue / linguistique de l'intuition vs. de l'observation

CORPUS = réservoir d'exemples non fabriqués

- Au sens technique qui prévaut aujourd'hui : un des objets même de la recherche

CORPUS = objet scientifique obéissant à des règles
et établi sur la base de principe

- Un corpus est un recueil de textes ou de paroles :
  - en format électronique
  - sélectionnés pour un objectif précis.

"A corpus is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language" (Sinclair, 1996)

- Le corpus pour la thèse en SDL : Définition de base : ensemble de **données langagières authentiques et attestées** (écrites et / ou orales) **organisées**, répondant à un objectif de recherche et remplissant **un certain nombre de conditions**. Ces données sont ensuite **préparées** pour donner lieu à des traitements automatiques plus ou moins poussés.

Vous êtes plutôt *based* ou *driven*?

- « (…) Corpus-based linguists adopt a 'confident' stand with respect to the relationship between theory and data in that they bring with them models of language and descriptions which the believe to be fundamentally adequate, they perceive and analyse the corpus through these categories and sieve the data accordingly. The corpus is considered useful because, on occasions, it indicates where minor corrections and adjustments can be made to the model and adopted and, of course, it can also be valuable as a source of quantitative evidence. » (Tognini Bonelli 2001 : 66)
  - utilisation du corpus postérieure à la formulation des hypothèses
  - rôle essentiel de vérification/validation

- « In a corpus driven approach the commitment of the linguist is to the integrity of the data as a whole, and descriptions aim to be comprehensive with respect to corpus evidence. The corpus, therefore, is seen as more than a repository of examples to back pre-existing theories or a probabilistic extension to an already well defined system. » (Tognini Bonelli 2001 : 85)
  - analyse du corpus antérieure à la formulation d'hypothèses
  - tout fait relevé doit être considéré comme pertinent
  - phénomènes absents aussi importants que phénomènes présents

Nouveaux usages, nouveaux objets de recherche, nouveaux corpus

- Les données **numériques natives** comme nouveaux corpus (Longhi 2012, Paveau 2013, 2015) => « écologie du discours numérique »
  - Nouveaux types de discours analysés / d'acteurs / d'interactions
  - Facilité d'accès trompeuse (droit, technique)
- Les **réseaux sociaux** comme nouveaux objets de recherche transdisciplinaire
  - Communication médiée par ordinateur (*CMC*) (Herring / Stein / Virtanen 2013)
  - Approche quali traditionnelle facilement doublée par quanti (Guilbert 2014, HS de *Corela*)

- Nouvelles formes d'**écriture** (Liénard 2011, 2012)
  - Poids des dispositifs socio-techniques sur les **pratiques** d'écriture:
    - Terminal
    - Émoticônes
    - Saisie intuitive
  - Poids de ces mêmes dispositifs sur le **résultat** de l'écriture (ici : le tweet)
  - Conséquences pour l'acte de décodage et de construction du sens

  ⇒**Brièveté** comme caractéristique clef : 140 caractères ici

  ⇒Plusieurs défis pour des études de **(micro-)linguistique empirique**

2. Problématique et objectifs

Les corpus de twitts : pour des approches intégrées

- Approche intégrée nécessite une **approche (micro-)linguistique** des tweets :
  - considérés dans leur face signifiante ;
  - envisagés comme 'micro-textes' (donc soumis aux / à des règles de **textualité**) ;
  - insérés dans un **dispositif** sociotechnique **interagissant** avec des formes de communication plus traditionnelles

Double problématique :

- Théorique :
  - Quels sont les **impacts du dispositif** sociotechnique sur la mise en œuvre des systèmes linguistiques considérés ?
  - Quelles conséquences doit-on en tirer pour l'appréhension de la **textualité du tweet** ?

⇒Questions testées ici à partir de écriture, opérateurs d'interaction et linéarisation

- Pratique :
  - Comment les scripteurs envisagent et gèrent-ils **la cohérence de leur dire** dans un cadre spatialement contraint et fonctionnellement prédéfini (opérateurs) ?

  => Quelle **littératie numérique** pour le « locuteur numérique » ?

« (…) la littératie numérique n'est **pas une catégorie technique** qui décrit un niveau fonctionnel minimal de compétences technologiques, mais plutôt une vaste capacité de **participer à une société qui utilise la technologie des communications numériques** dans les milieux de travail, au gouvernement, en éducation, dans les domaines culturels, **dans les espaces civiques,** dans les foyers et dans les loisirs ». (Hoechsmann / DeWaard 2015 : 5)

3. Quelles données ? Le défi de la constitution du corpus

De nouvelles pratiques de collecte

- Dimensions juridique et éthique :

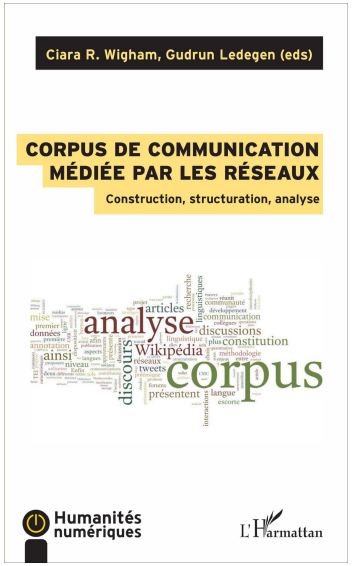
Twitter ne revendique aucun droit de propriété intellectuelle sur les contenus produits par les utilisateurs du service. (…) Mieux encore, Twitter encourage ses utilisateurs à verser les contenus par anticipation dans le domaine public ou à les placer sous licences libres pour en favoriser la réutilisation. (Blog SI Lex de Lionel Maurel)

- Dimension technologique : compilation des données *via* l'API de twitter
- Dimension « archivistique » : gestion des métadonnées, structuration (TEI)

L'indispensable phase d'annotation

- Indexation des métadonnées pour la gestion des interactions

2	LaflotteJohan||Johan Lafiotte||RT @david_pellard: Une pensÃ©e pour @sandrinebeller dont l'excellent travail au parlement europÃ©en n'a pas rÃ©sistÃ© au choc du #fn #EP2014 #â€¦|||
3	SamuelOguenine||Sam OguÃ©nine||RT @MarieBerrube: Par contre, ceux qui twittent "la France aux Arabes" j'crois qu'ils se sont trompÃ©s aussi. Ok on est contre le FN mais y'â€¦|||
4	Benjie_Tonitrue||Benjie_Tonitrue||RT @D1Frc: #Hollande : "Il faut regarder la victoire du FN en face". Traduction : Pas de soirÃ©e au thÃ©Ãtre du Rond-Point pour cette fois. â€¦|||
5	PasRateurFTV||PasRateurFTV||RT @CraissacLJM: PETIT CALCUL : 25% de 43% de votants = 4,75millions d'Ã©lecteurs pour le #FN. Au 1er tour de la prÃ©sidentielle 2012 : 6millâ€¦|||
6	SavalleIsaline||Isaline.||RT @CeliaJankowski: Je sais pas qui me dÃ©goÃ»te le plus, ceux qui votent FN ou ceux qui se plaignent que le FN gagnent alors qu'ils ont pas â€¦|||
7	Mou2s_||Corleone||RT @mohadehilis: Le FN fait d'la Pen! Punchlinnnnne||
8	BFMTV||BFMTV||EN DIRECT - #Europeennes2014 : Hollande rÃ©unit lundi Ã 8h30 Valls et plusieurs ministres #EP2014 http://t.co/0c0qMqNgn6||
9	Melka_L||ÙfØ§Ù…Ù„||Le FN il on fais vla le score||
10	Elle_ize||Elise ZÃ©lie ||RT @LinksTheSun: Le FN est en progression. Je suis aussi surpris que le jour oÃ¹ j'ai appris que Ricky Martin Ã©tait homosexuel.||
11	Djidji_Sahar||Ø§Ù„Ù‡ ØµÙfØ¨Ø± â˜ª||RT @AlassaneDashiki: Vous vouliez pas du FN bh fallait votÃ© mon frÃ¨re||
12	aisman69||angel ibaÃ±ez||RT @LePoint: #LePointLive La carte des rÃ©sultats en France. En marron, les victoires du #FN. #EP2014 http://t.co/jRgFIZNUeU||
13	FarahGday||â˜†tounsiaâ˜†||RT @AlexisPTR: Voter FN sans regarder leur plan politique juste parce que vous vous Ãªtes fait volÃ© votre 5S dans le RER par 2 reubeu c pas â€¦|||
14	A2Politique||Pierre||RT @BFMTV: EN DIRECT - Les conservateurs revendiquent la prÃ©sidence de la Commission europÃ©enne http://t.co/0c0qMqNgn6 #EP2014||
15	lysandre87||lysandre merlier.||RT @Le_Ptrolls: Guaino a votÃ© FN ? #JePoseLaQuestion http://t.co/aGxaEvKsGG||
16	brynlhmm||WHERE IS BRYAN ?||RT @lemondefr: PoussÃ©e du FN dans toutes les circonscriptions http://t.co/7FQlJb0Ukt||
17	LilyCathyT||Lily.||RT @Tilahh: Le FN Ã 25%. Profitez bien de vos derniers moments de libertÃ©, on revient bientÃ´t dans les annÃ©es 40.||
18	JulesMtp||Ì°||RT @NicolasSch : Les votants FN, je veux pas vous entendre vous plaindre quand le kebab en bas de chez vous sera remplacÃ© par un bar Ã salaâ€¦|||



- Annotation du contenu textuel => défi méthodo pour opérateurs, gestion de l'orthographe + tagging espèces de mots + analyse syntaxique (*cf. infra*)

Un corpus original : tee2014

- MSH Dijon (TIL, Cimeos, LE2i) + Le Havre + Metz + partenaires dans 4 pays européens => 5 terrains nationaux
- Objet : communication « générée » par les candidats aux Elections Européennes de 2014 => 80 comptes par pays
  - Les messages envoyés sur les comptes Twitter des candidats
  - Les messages inclus dans les « conversations » entre ces comptes et d'autres tweetos (discours citoyens, débats internes…) ;
  - Les messages contenant les "hashtags" sélectionnés, liés à des thématiques politiques majeures de chaque pays
- 4 semaines de collecte : avant et juste après le scrutin

Extraction du corpus global sur 2 langues

- Corpus « français » et « allemands » => liés aux comptes des candidats français et allemands (même si hétérogénéité linguistique)
  - F : Plus de 1 millions de tweets
  - D : 720.000 tweets

  ⇒Toujours RT compris
- Traitement pour interrogation (semi-)automatique
  - Deux sorties : aspiration complète avec méta-données + texte brut
  - Interrogation sous AntConc (passage dans TXM en cours)

Forces et faiblesses

- Analyse du texte 'brut' : un tweet = un texte, avec insertion possible dans des séries d'interactions (cf. infra)
- Pas de procédure on-/off-line de saisie des stratégies de production/réception
- seule voie d'accès : tout ce qui relève du métalinguistique + opérateurs # @ RT http (cf. infra)
- Cohérence dans les tweets (politique (de campagne électorale))

4. Quels cadres méthodologiques ?

Les DH ne font pas table rase du passé. Elles s'appuient, au contraire, sur l'ensemble des paradigmes, savoir-faire et connaissances propres à ces disciplines, tout en mobilisant les outils et les perspectives singulières du champ du numérique. (Art. 2)

Linguistique des médias ?

- Une tradition largement germanique et scandinave :
- « Linguistique des médias » (Burger, Perrin, Hauser, Lingenbühl) incluant la multimodalité (Held, Stöckl) et interrogeant ses relations avec les SIC (Bucher, Schmitz)
- Mise en perspective des approches linguistiques et communicatives entre
  - une focalisation des SIC sur le niveau englobant
  - Relation émetteur – récepteur (intention) – réseaux
  - Cadre institutionnel du processus de communication

- Des phénomènes de niveau microlinguistique en lien avec les caractéristiques de l'acte de communication global :
  - dans une perspective top-down : comment la situation de communication se répercute-t-elle sur certains choix langagiers ?
  - dans une perspective bottom-up : comment certains traits langagiers assurent-ils la réussite de l'acte de communication dans la situation (contrainte) donnée ?
- => Twitter semble représenter un cas d'école

Analyse de discours linguistique et outillée

- Analyse linguistique du discours politique
  - F : tributaire de l'héritage de l'Ecole Française d'Analyse du Discours (Mazière 2005)
  - GB : *CDA* (Semen 27 = Schepens/Petitclerc 2009)
  - A : Sémantique discursive (historique) (Busse 1987,1997, Stötzel/Wengeler 1995, Spitzmüller/Warnke 2011, Busse/Teubert 2013)
- Analyse linguistique du discours médiatique
  - F : née de l'analyse du discours (Charaudeau 1997, Moirand 2007), focalisée sur la presse et les médias-audiovisuels
  - GB/A : autonomisation d'une 'linguistique des médias' *cf. supra*

Textualité comme clef d'entrée

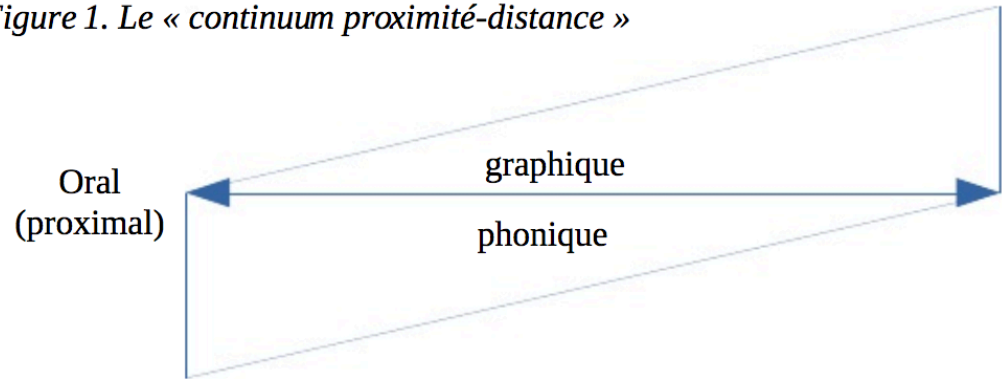
- Quelle est la validité de la notion pour la communication électronique ?
- Comment la notion de textualité s'articule-t-elle avec les normes communicatives ?
- Point de départ : définition classique en ling. text. : ce qui fait d'une suite de signes un texte
- Notions clefs : cohérence, cohésion, … (De Beaugrande / Dressler 1981)

5. Discussion

Ecrit ou oral ? Plutôt distance ou proximité…

- Un problème mal posé
- Saisie du continuum écrit/oral (Koch/Österreicher 2011, présentation en français *cf.* Modicom 2015) reposant sur l'opposition entre :
  - Niveau du média
  - Niveau conceptionnel

Figure 1. Le « continuum proximité-distance »



Modicom (2015 : 4)

(K&O 2008:201 ; version plus complexe chez K&O 1985:23, et ci-dessous en annexe I)

- Interactions «orales » analysées à replacer le long d'un continuum « proximité communicationnelle » vs. « distance communicationnelle »
- Paramètres essentiels :
  - Interaction publique ou privée
  - Degré de familiarité
  - Part laissée aux émotions
  - Intégration de l'action et de la situation
  - Co-construction de la référence
  - Proximité physique
  - Coopération entre interlocuteurs
  - Dialogicité
  - Spontanéité limitée
  - Choix de l'objet de discours imposé ou non par l'interaction

- (1) .@ShaeCald1 **l**es médias usent de la langue de bois. Ils font eux aussi du populisme! La vérité c'est que **+ 10%** des français sont **cons**! #FN
- (2) RT @Moha212_: RTsi rebeu avec ta meuf babtou -Ali faut qu'on casse -mais pk bébé? -le FN est passé tu va retourner au bled ciao -mais :(
- (3) RT @CecileDuflot: Le budget de la campagne d'@EvaJoly c'était 1,7 million d'euros. Pour toute la campagne. Oui. Toute. #deladémocratie #com…

- Réduction du degré d'informativité **explicite** à un minimum :
  - En Conseil Municipal à la mairie du 9ème
  - Mir etwas zu sehr auf "arme #afd-" gepolt, aber einige richtige, wichtige Ansätze dabei in der @SZ.

  => Déplacement du **lieu d'inscription de l'informativité**
- Nécessité d'injecter dans le décodage outre la situationalité de départ le savoir fonctionnel lié aux opérateurs @, http:// et #

  => Vers une cohérence segmentée

Le corpus « twitter » ou les interactions revisitées

Les opérateurs techniques de Twitter comme marqueurs d'opération discursives (Thimm/Dang-Anh/Einspänner 2012)

Operators	Text	Function
@ (addressing, mentioning)	Name of account, locational reference, emoticons, part of e-mail address	Personal reference, creating attention, dialogicity/interaction, contacting, responsivity, intertextuality, coherence
# (indexing)	Lexeme, key word, occasionalism, abbreviation, acronym, cumulated phrases	Contextualization (cue), discourse organization, topical referencing, tagging, creating ad hoc publics, rhetorical branding
http:// (hyperlinking)	Other websites: internal and external links, pictures/photos (e.g. Twitpics), videos (e.g. YouTube)	Information diffusion, argumentation, illustration, quasi evidencing
RT (Retweet) (redistributing)	Citation, comment	Diffusion, reference, citation, creating attention

Cohérence interactionnelle

- Enjeu pour l'interprétation des opérateurs @ et RT
- RT comme marqueur de diffusion, mais quid de la prise en charge énonciative surtout avec polyphonie exponentielle ?

(4) RT @LCI: @Herve_Morin sur #UMP : "La période va être terrible. Elle amènera à une recomposition" #LCI @ACrespoMara

(5) RT @FelixMSteiner: "Nicht #wählen die #AfD du darfst!" Danke @_verdi dafür! #Europawahl #Humor http://t.co/ogA6EkBang

- Double statut de @ comme marqueur d'adresse (9) et/ou mention (10) :

(6) Merci @TomChevalier76 @Parsquiou @Pilouilleuh @Dragon76000 Thibaut Vs êtes la jeunesse européenne ! @UDIjeunes76 http://t.co/m4djWbScAX

=> Décodage dépendant du remplissage textuel du tweet : *vous* d'adresse, acte de langage exclamatif + fonction phatique de *merci*

(7) Ich lach mich tot."#AfD - Auffangbecken für Dumme" Danke an @EinAugenschmaus für diese herrlich treffende Umbezeichnung.

=> Triple statut en contexte : mention + source + marqueur d'adresse

Linéarisation, Cohérence et opérateurs

- Remise en cause des principes de linéarisation comme base de construction de la cohérence :
- (8) RT @david_attia: En Conseil Municipal à la mairie du 9ème arr. @dburkli @jbdef @JuliaSereni @MkFAURE #TeamBurkli http://t.co/kEKFkTOB4F
- (9) RT @PorcusDivinus: Mir etwas zu sehr auf "arme #afd-" gepolt, aber einige richtige, wichtige Ansätze dabei in der @SZ. http://t.co/3K14pHF…
- => Enjeu : garantir l'acceptabilité du tweet comme texte

et renégociation de la cohérence

- Double statut de #
- # intégré syntaxiquement

(10) RT @Cesarmand: @DominiqueRiquet élu @Europarl_FR VP Commission #Transports parle #GrandParis et #Interconnexions https://t.co/jxa4N0cC5W #E‰Û_

(11) @OlgaVanHorst @DMWarrior @BiboBissig Was konkret ist im #afd Wahlprogramm lächerlich? Ich finde es sehr fundiert!

  - Décodé comme arguments de l'énoncé (« niveau 0 ») = emploi mondain
  - Décodé comme tête de paradigme = emploi autonymique

- # non intégré syntaxiquement, en position d'ouverture (rare) et/ou de clôture (en masse)

(12) #Lille: "On a accusé l'Europe de ce qu'elle ne pouvait pas faire. L'Europe est 1 champion économique et démocratique " #LesEuropeens cc

(13) #AfD Stand vor dem Wittwer! Polizei hat bereits den gesamten #Schillerplatz umstellt! #stuttgart #bunt statt #braun #wk14 #reclaimeurope

- Fonction instructionnelle, « cadrative » impliquant/nécessitant un décodage segmenté
- Participe à une cohérence de double niveau
  - Interne au texte
  - Supérieur au niveau du fil de tweets

Les # comme forme ultra-brève de problématisation et d'orientation du lecteur ?

- Mise en œuvre d'un trait du dispositif sociotechnique de Twitter (*cf. infra*) : le **#**
  - Quelles fonctions ?
  - Quelle instrumentalisation possible ?
  - Démultiplicateur ou réducteur de brièveté ?
- Etude de cas en **communication politique**
  - Le # comme nouvelle forme de « petite phrase » ?
  - Le # comme indice de « mise en formule » ?

- Forme d'indexation permettant, selon la position du #
  - de cadrer le message avec possible topicalisation thématique

(14) **#Bruxelles** ...il etait fan de Dieudonné et de la Quenelle comme JMLP et Gollnisch?

(15) **#Musée #juif de #Bruxelles** : un Français soupçonné d'être le tireur arrêté à Marseille - http://t.co/9XgP6iia8V via @LorientLeJour

(16) **#Berlindirekt** Frau von der #Leyen, warum redet niemand mit #Putin ? Warum kann SchrÌ¦der und Frau von der Leyen nicht ?

(17) **#Europawahl** was tun gegen Islamfeindlichkeit? Islamophobe und islamfeindliche Einstellungen, wonach der Islam... http://t.co/KEFMfFdNs8

– De résumer le message

(18) Copé a demandé aux Français de sanctionner Hollande aux européennes. Ils l'ont écouté... En votant FN. **#bravo #bienjoué**

(19) "RT @marteria: Morgen aufstehen und wählen gehen !!! #EP2014 #Europa **#GehtAuchAnders** http://t.co/Oj6wwrpZRR"

– De l'inscrire dans un flux (**+** formule)

(20) Le Pen ? ça sent l'gaz ! #Coupdecrayon de Troud #FN #Haine **#laluttecontinue** #electioneuropeennes http://t.co/MAxocABEJo

(21a) **#unautreregard** *#NotreEurope* RT@UnionAgricole87: #reportageMultimedia Le meeting européen de @jpdenanot @PS_EP2014 http://t.co/qaYY6woDPV

(21b) @tjoubert101 @jpdenanot @PS_EP2014 Merci pour le hashtag **#unautreregard**. J'apprécie.

(22) RT @OpenEurope: .@MartinSchulz: The pan-European candidate? #EP2014 #EU #Bild #Europedecides **#Schulz #jetztistschulz #nowschulz** http://t.co‰Û_

6. Perspectives

- Tweet comme objet d'étude à part entière et extrêmement complexe pour les sciences du langage :
  - différence qualitative avec d'autres formes d'écriture électronique (FB, blogs, forums)
  - enjeu épistémologique : construction d'une « théorie » du signe tweet revisitant des catégories pourtant anciennes : textualité, écrit/oral, recevabilité…
  - enjeu méthodologique : analyse qualitative devant inclure le balayage quantitative de grandes masses

- Objet de recherche fécond pour l'interdisciplinarité (sciences du langage, info-com, science politique, sciences de l'éducation…)
- Limites claires des approches « décontextualisées » :
  - Production de # : choix, intentions
  - Réception des # : nouvelle syntaxe d'écriture nécessitant de nouvelles habitudes de lecture / part de « remplissage » sémantique individuelle importante
  - Pour des # « viraux » : nécessité d'une analyse automatique des flux
- Vrai enjeu : apparition de nouvelles formes de littératie numérique comme les # nécessite une vraie appropriation critique du dispositif + nouvelle « gestion » cognitive de la brièveté

Merci pour votre attention !

Laurent Gautier

Université Bourgogne Franche-Comté (EA4182)

laurent.gautier@ubfc.fr