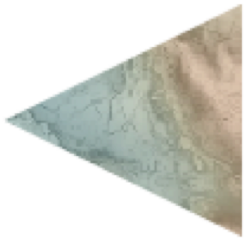
De la multiplicité des données en langue naturelle aux corpus : contraintes et possibilités méthodologiques (en discours spécialisés)

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



Structure

1. Les décisions incontournables
2. Constituer « son » corpus
3. Les corpus pour les travaux contrastifs
4. Trois focus sur des corpus constitués en LSP
5. 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

- Conception **minimaliste** du corpus :
  - représentativité par rapport à quelle norme ?
  - biais de sélection par le chercheur ?

La linguistique ? de corpus…

- Peu de domaines (aucun ?) des modalités de communication résistent encore à la « mise en corpus » : interactions orales (audio + vidéo), échanges multimodaux, LSF (traces)
- Au sens technique qui prévaut aujourd'hui : un des objets même de la recherche
- Héritage des travaux en analyse conversationnelle depuis l'ethnométhodologie des années 1970

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.

Un impératif pour toute approche cognitive

- Lecture « à rebours » des 4 grands principes de la linguistique cognitive (Geeraerts 2006) et de leurs liens avec les corpus :
- Le sens linguistique repose sur **l'emploi et l'expérience** : l'emploi est donc révélateur d'un certain mode de conceptualisation => programme de Wittgenstein ?
  - Plaidoyer pour l'analyse de larges corpus écrits et oraux
  - Combinaison entre études sur corpus (corpus-based) et de corpus (corpus-driven)
- Le sens linguistique a une **dimension encyclopédique** et **hétéronome** :
  - Corpus comme « réservoir » des dimensions enfouies : exemples des *default values* dans la sémantique des cadres

Un impératif pour toute approche cognitive

- Le sens linguistique est **dynamique et flexible** :
  - Corpus comme ressource constructiviste : traces de la construction du sens (vs. objectivisme)

- Le sens linguistique est une **mise en perspective du monde** (re-examen de l'hypothèse de Sapir-Whorf)
  - Nécessité de prendre en compte l'environnement du corpus, en termes de *ecology of language* (Haugen)
  - Conditions de production cruciales :
    - Corpus pré-existant au travail de recherche : opération de sélection du linguiste
    - Corpus expérimental : produit pour le chercheur

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

2. Constituer « son » corpus

Conditions essentielles

- En **adéquation totale** avec l'objet linguistique de la recherche
- Critère d'**authenticité** : donné la plupart du temps, mais à traiter à précaution quand le corpus doit être recueilli

- Critère d’**homogénéité** multi-niveaux
  - temporelle
  - spatiale
  - thématique
  - énonciative

  ⇒Poids des différents niveaux dépend de l’objet de la recherche
- Critère de **représentativité**, garant de la validité des résultats

=> Question de comparabilité et, donc, de corpus de référence / étalon

Critères de sélection des textes

- typologie des textes, genres et registres (D.Biber, 1999)
  - types de textes: ensemble de caractéristiques linguistiques
  - genre/registre: catégories intuitives qui utilisent les locuteurs pour répartir les productions langagières
    - thèmes
    - formes de textes

- Paramètres situationnels (Biber 1999)
  - Canal : écrit/parlé/écrit lu;
  - Format : publié/non-publié;
  - Cadre : institutionnel/autre cadre public/ privé-interpersonnel
  - Destinataire :
    - Pluralité : pluriel/individuel/non-compté
    - Présence : présent/absent
    - Interaction : aucune/peu/beaucoup
    - connaissances partagées : générales/spécialisées/personnelles

3. Les corpus pour les travaux contrastifs

- Deux principaux types de corpus (indépendamment de la perspective choisie) en approche contrastive :
  - corpus parallèles
  - corpus comparables (terminologie d'après Teubert 1996, corpus 'multilingues' chez Baker)
  - troisième type postulé par Baker : corpus 'comparables'

Une opposition tellement classique qu'elle s'en est fossilisée : Teubert (1996)

- « A 'parallel corpus' is a bilingual or multilingual corpus that contains one set of texts in two or more languages. » (Teubert 1996 : 245)
  - **Textes traduits / bi-textes**
  - *Tertium comparationis* : équivalence supposée entre les textes
  - préparation nécessitant une opération d'alignement (semi-)automatique
  - Exemple : conférence de presse de la BCE (8/an) : TS anglais vs. TC dans les 23 langues

Anglais	Français	Allemand	Néerlandais
Given high structural unemployment and low potential output growth in the euro area, a cyclical recovery along the lines of the March ECB staff projections is no grounds for complacency.	Au vu du niveau élevé de chômage structurel et de la faible croissance potentielle dans la zone euro, une reprise conjoncturelle telle que celle ressortant des projections de mars des services de la BCE **ne permet aucun excès de confiance.**	Angesichts der hohen strukturellen Arbeitslosigkeit und des geringen Wachstums des Produktionspotenzials im Eurogebiet **gibt** eine Konjunkturerholung wie in den von Experten der EZB erstellten Projektionen vom März **keinen Anlass zur Sorglosigkeit.**	Gezien de hoge structurele werkloosheid en de lage potentiële productiegroei in het eurogebied, **is** een conjunctuurgebonden herstel zoals geschetst in de door medewerkers van de ECB opgestelde projecties van maart **geen reden om achterover te leunen.**

Quels niveaux d'**équivalence** postulés (aveuglément ?) pour la comparaison ? :

-dénotative
-conntotative
-pragmatique
-textuelle
-stylistique/esthétique (Koller 1992)

- Utilisations principales :
  - recherche d'**équivalents** terminologiques (*Weltwirtschaft/économie mondiale)* : domaine plus ou moins autonome en terminologie (ATR) / Het herstel van de binnenlandse vraag
  - recherche de **collocations** *(la croissance du PIB réel/das Wachstum des realen Bruttoinlandsprodukts/de reële bbp-groei*)
  - recherche de traits **stylistiques** ? *était supérieur à la tendance/übertraf den Trend*
- « Inconvénients »/risques :
  - tributaire de la **finesse de l'alignement**
  - tributaire de la **qualité** de la traduction
  - quelle valeur pour les résultats en dehors de l'opération de traduction ? Cf. travaux de l'UMIST (M. Baker) sur le « style du traducteur »

- « 'Comparable corpora' are corpora in two or more languages with the same or similar composition. All copora have an explicit or implicit composition. The texts they contain can be classified according to a variety of intralinguistic or extralinguistic features. » (Teubert 1996 : 245) :
  - Tertium comparationis : à définir en fonction des objectifs à atteindre car seul **garant de l'homogénéité** du corpus constitué
  - ⟶ un tc inscrit au niveau textuel / discursif
- TC possibles :
  - un type de texte
  - un domaine thématique
  - une situation énonciative
- Apports essentiels :
  - analyse de deux langues « originales »
  - travail aux niveaux textuel et discursif
  - travail possible au niveau microlinguistique (ATR par exemple) mais sans appariement automatique

- « Inconvénients »/risques :
  - précision dans la définition du tc (elle-même tributaire du niveau d'analyse)
  - part d'idiosyncrasies « culturelles » (danger pour les appariements automatisés)
  - part importante des choix initiaux et des a priori (exemple des types de texte)

Un corpus complexe

- Conférences de presse de la BCE traduites
  - Textes compilés depuis le site de la BCE - https://www.ecb.europa.eu
  - 8 interventions en 2015 et 8 en 2016
  - Extraction, alignement et analyse des textes en anglais, français, allemand et néerlandais
- Rapports trimestriels de la Banque Nationale Suisse
  - Textes compilés depuis le site de BNS - https://www.snb.ch/fr/iabout/pub
  - 4 par ans Extraction, alignement et analyse des textes en français et allemand
- Perspectives économiques de la Banque nationale de Belgique
  - Textes compilés depuis le site de BNB - https://www.nbb.be/fr/publications-et-recherche
  - 2 par an
  - Extraction, alignement et analyse des textes en français et néerlandais
- Bulletins de la Banque de France
  - Textes compilés depuis le site de la BdF https://publications.banque-france.fr/liste-chronologique/le-bulletin-de-la-banque-de-france + anglais
  - 6 par an

Un réseau de relations à démêler

- BCE : anglais langue originale + allemand / néerlandais / français traduit

⇒ double corpus comparable nécessaire : Banque d'Angleterre, *Bundesbank*, Banque de France

- Banque Nationale Suisse : co-rédaction dans les 3 langues + anglais traduit ?

⇒ triple corpus comparable nécessaire : Banques nationales allemande, italienne et BdF

- Banque nationale de Belgique : co-rédaction dans les 2 langues + anglais traduit ?

⇒ Double corpus comparable nécessaire : Banque de France et Banque néerlandaise

- Banque de France : français langue originale + anglais traduit

=> Quel traitement du plurilinguisme ?

4. Trois focus sur des corpus constitués en LSP

Les corpus en LSP : des terrains sensibles ?

- Rôle clef de ce qui est « autour » du corpus => approche ethnographique / écologique du corpus (Gautier 2019a)

Fragments of discourse materials always are shaped and constrained by the larger organizational settings in which they emerge and simultaneously influenced by cognitive / emotional processes despite the convenience of only focusing on extracted fragments independently of the organizational and cognitive/ emotional complexity of daily life settings. (Cicourel 2007 : 736)

- Un nécessaire aller-retour entre approches quantitatives et qualitatives, non exclusives

Focus 1 : discours politique sur twitter
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**

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)

De nouvelles pratiques de collecte

- Dimensions juridique et éthique :

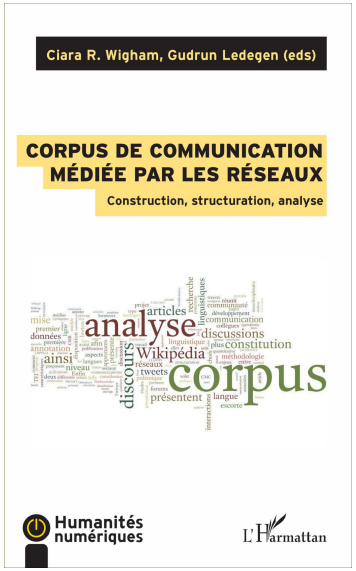
Twitter ne revendique aucun droit de propriété intellectuelle sur les contenus produits par les utilisateurs du service. (…) Mieux encore, Twitter encourage ses utilisateurs à verser les contenus par anticipation dans le domaine public ou à les placer sous licences libres pour en favoriser la réutilisation. (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	LafotteJohan||Johan Lafotte||RT @david_bellard: 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©Ã¢tre 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||ÙƒØ§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„Ù‡ Ø£ÙƒÙ¨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||Î”||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 (bientôt… tee2019)

- 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))

- (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

Focus 2 : discours sensoriel autour du cacao

- Nécessité d'un **décentrage « culturel » et méthodologique**
  Focus sur un pays producteur, ici : l'Equateur
  – Saisie *in vivo*

Cabosses de cacao, contient 15 à 40 « fèves »

Chocolat pressé entre des feuilles de « Wijao »

Coopérative de producteurs Chocounión – collaboration entre la Agencia de Cooperación Española et la Pontificia Universidad Católica del Ecuador Sede Esmeraldas

L'approche « ethnographique » des LSP comme préalable

- Prolongement de l'approche des champs spécialisés :
  - Triangulation de langue, discours et culture (*textography,* Swales 1998)
  - Saisie de l'environnement spécifique dans lequel se construisent les discours spécialisés (revue chez Dressen-Hammouda 2013)
- Mise en œuvre de méthodologies ethnographiques (*using ethnographic tools*, Heat / Street 2008 : 120) en amont de la saisie des discours et donc de la construction des corpus (Wozniak 2011, 2012 Resche 2013, Isani 2014):
  - Immersions *in situ*
  - Observation => Notes de terrain
  - Questionnaires
  - Entretiens semi-directifs, interviews longues
  - Revues professionnelles

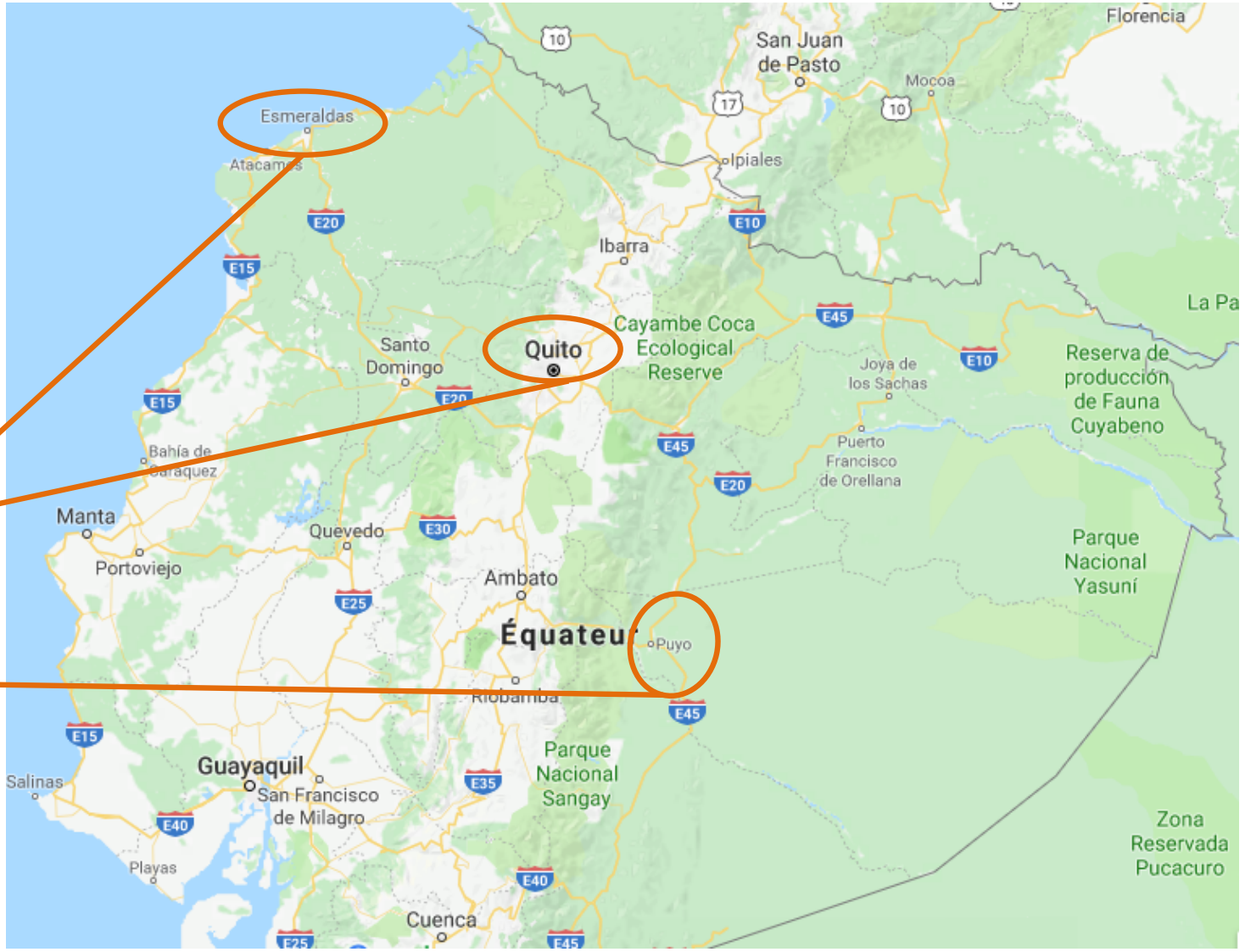
La vinculación también implica la asistencia técnica a los **indígenas** en el manejo del **cacao**, como una opción para mejorar sus condiciones económicas.

A través de la vinculación se quiere motivar a los jóvenes **chachis** para que aprovechen las tierras de la **nacionalidad** y sean **cultivadas** técnicamente y con orientación profesional.

Zone d’investigation de Javier

Capitale

Zone d’investigation de Leticia & Olivier

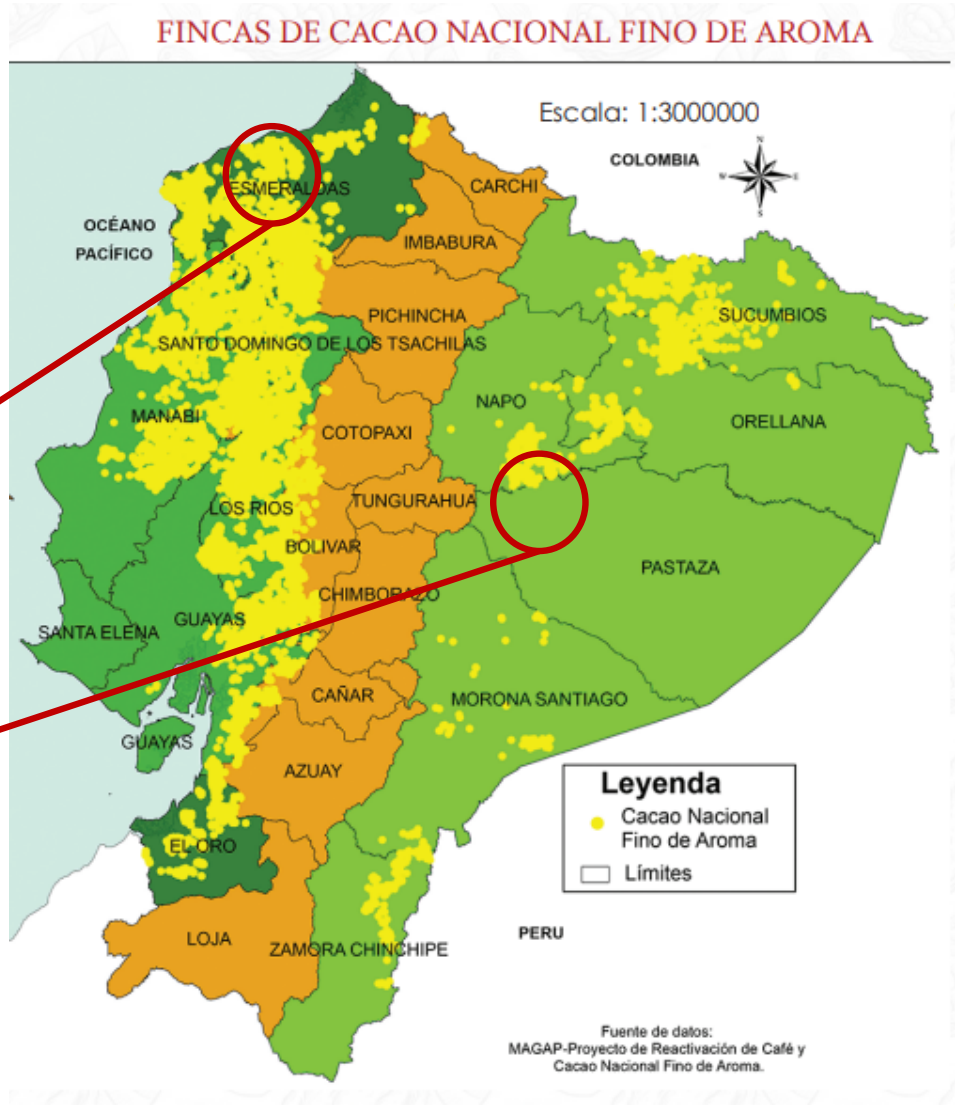


Fuente: TERCER SEMINARIO ANUAL DE CACAO EN LAS AMÉRICAS
6 DE SEPTIEMBRE 2016, GUAYAQUIL, ECUADOR
http://www.worldcocoafoundation.org/wp-content/uploads/files_mf/1474315649ELLIBRO.pdf

Secteur de production oú se trouve Javier

Depuis 2014-2016 de nouvelles zones sont apparues oú nous allons faire les enquêtes

Liens vers plus d'informations socio-économiques:
http://sipa.agricultura.gob.ec/index.php/cacao
http://www.anecacao.com/index.php/en/estadisticas/estadisticas-actuales.html
http://visit.ecuador.travel/chocolate/



Une approche socio-linguistique

- Approche « naïve » : lexique espagnol
  - **Terminologie commerciale/marketing** diffusée, dans le cas particulier du cacao en Amérique Latine, par les *aficionados*

  => Quel positionnement par rapport à catégorie traditionnellement mobilisée de « prescripteurs » ?

- Approche *in vivo* : nécessité de tenir compte des langues natives :
  - le cha'palaa (famille barbacoanne), langue vernaculaire de la communauté chachi spécialisée dans les productions agricoles (cacao, coco, bananes),
  - le kichwa (famille Quechua IIB), langue vernaculaire de différentes communautés amazoniennes

  => **Terminologie de filière**

Des méthodologies à inventer

- Nécessité de recourir à des **données**
  - **Attestées** en fonction des **usages** à étudier (

 aux compilations de glossaires, dictionnaires, métaglossaires !)
  - **Authentiques** = produites originalement dans chaque langue interrogée (=> corpus comparables, Teubert 1996)
  - Produites si nécessaire **expérimentalement** (Gautier/Hohota 2014, Bach 2017, Mancebo-Humbert/Le Fur/Gautier 2018, Gautier 2018)
  - **Annotées** et **structurées** permettant un traitement quantitatif en plus de qualitatif pour **éviter la fétichisation d'hapax**

Corpus 1 : « discours ambiant » et marketing (Gautier 2014, Gautier *et al*. 2015) :

- Sorte de corpus « témoin »
- Identifier les descripteurs spécifiques au discours sensoriel utilisé en espagnol pour décrire le chocolat en contexte équatorien `

- Discours prescriptifs / des prescripteurs :
  - Ensemble des discours « de ceux qui comptent » dans la filière / argument d'autorité
  - Prototypes : grandes revues, guides reconnus, dégustation de formation ou de concours (Gautier / Hohota 2014)
  - Fonction essentiellement évaluative puis directive (classement, incitation à l'achat)
  - Rôle important dans la diffusion des termes et des concepts

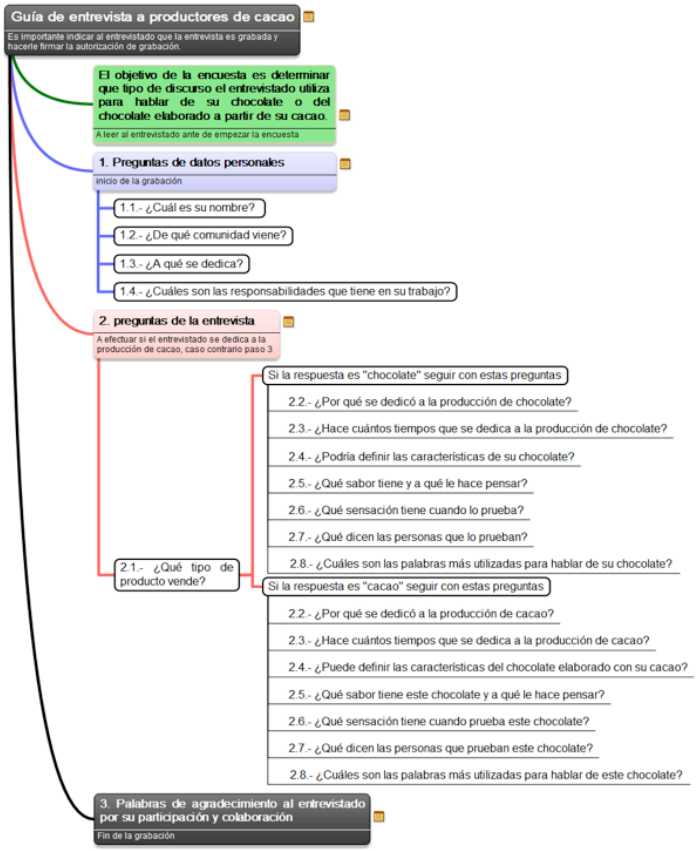
- Discours descriptifs / évaluatifs :
  - Ensemble « flou » (fuzzy) de discours ayant le cacao comme objet mais ayant des sources « non autoritaires » (pb. description/définition du non-expert/amateur)
  - Prototype : revues semi-spécialisées, pages « critique œnologique » des quotidiens, blogs d'amateurs, dégustation 'marketing' (Gautier/Hohota 2014)
  - Fonction : évaluative
- Discours marketing et publicitaire :
  - Ensemble lui aussi « flou » avec apparition de formes hybrides (publi-communiqué…) avec dimension commerciale comme dénominateur commun
  - Prototype : packaging, publicités
  - Fonction : directive (achat), évaluative et expressive

Corpus 2 : corpus d'entrevues / entretiens avec les producteurs

- Objectif : interroger leur mode de mise en discours des descripteurs précédemment identifiés ;
- remonter aux descripteurs premiers dans les deux langues natives mentionnées
- Tenir compte des contextes locaux :
  - producteur vs.
  - producteur-transformateur (différence vin)

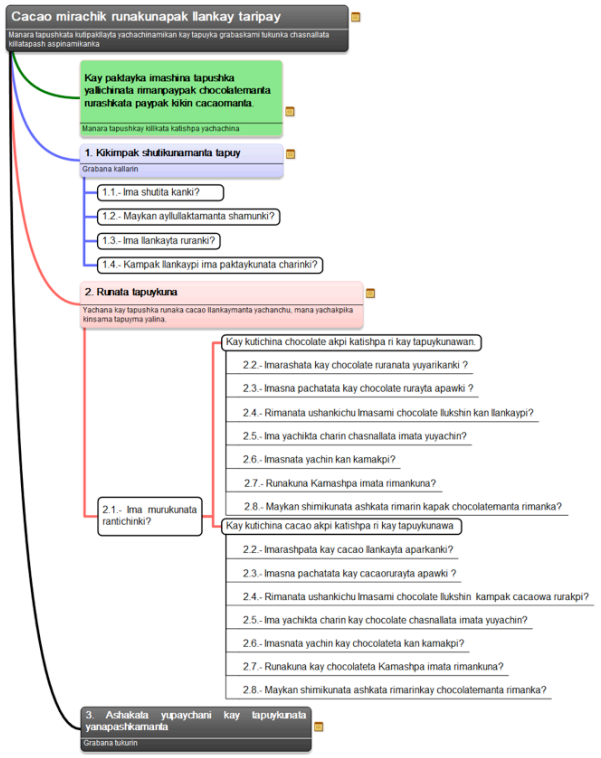
Guía de entrevista a productores de cacao

Es importante indicar al entrevistado que la entrevista es grabada y hacerle firmar la autorización de grabación.

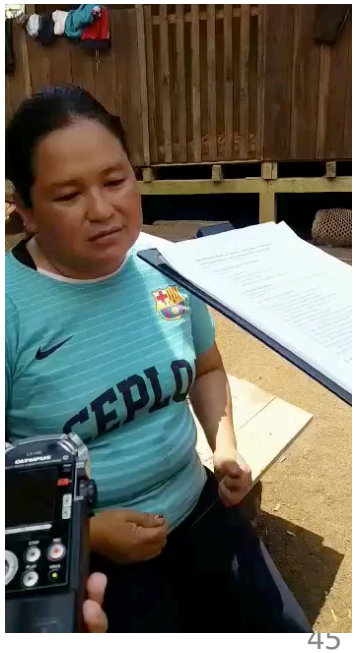


Cacao mirachik runakunapak llankay taripay

Manara tapushkata kutipakllayta yachachinamikan kay tapuyka grabaskami tukunka chasnallata killatapash aspinamikanka.



Exemple en Kichwa (Amazonie Equatorienne)



Corpus 3 : un corpus mixte de ‘dégustation’

- questionnaires sans stimulus visant à identifier la valeur sémantique des descripteurs à partir d'un processus progressif d’abstraction progressive (Dubois 1995)

- transcriptions issues de la dégustation d’un chocolat choisi pour être représentatif de chaque descripteur.

Focus 3 : Les corpus expérimentaux dans le sensoriel

- Deux types de corpus expérimentaux : avec et sans stimulus.

Les corpus **sans stimulus**

- Deux objets d'étude, deux corpus :
  - le corpus *Questionnaire Crémant de Bourgogne*
  - le corpus *Questionnaire espumantes*

  - consistent à faire répondre des ***consommateurs*** à des questions sur les produits en question, sans référent, pour une génération de paroles en dehors de toute situation de dégustation, basée sur les connaissances ou les expériences antérieures des individus, la mémoire expérientielle.
  - permettent d'approcher les modes de construction sémantique des termes descripteurs en observant les stratégies de mise en discours (hésitations, reformulations, explications) et la construction des prototypes.

- Approche dite « de terrain », empirico-inductive (Blanchet, 2012)
- Questionnaire unique :
  - pour les deux objets : Crémant de Bourgogne et *espumante*
  - dans les deux situations de collecte : sans et avec stimulus
- L'élaboration du questionnaire répond aux objectifs de l'étude et prévoit une analyse linguistique sémantico-cognitive (Mondada, 1998) :
  - un questionnaire semi-directif : des questions ouvertes

permettant « d'identifier les représentations cognitives, en particulier en vue de déterminer les propriétés sémantiques d'un concept […] et la manière dont on peut identifier leurs relations à des catégories cognitives. » (Delepaut, 2009 : 164)

- Questionnaire composé de 3 questions complémentaires => trois sous-corpus de parole.

Q1. Si je vous dis « Crémant de Bourgogne », à quoi pensez-vous ?
- sous-corpus « évocation » : qui permet de verbaliser de façon spontanée et intuitive l'image du Crémant de Bourgogne présente à l'esprit des répondants.
Q2. Si vous deviez expliquer à un ami ce qu'est un Crémant de Bourgogne, que lui diriez-vous ?
- sous-corpus « explication » : qui permet d'accéder à un niveau d'abstraction plus élevé incitant au dépassement de l'évocation intuitive pour aborder des aspects « techniques ».
Q3. Si vous deviez choisir trois mots ou expressions pour définir le Crémant de Bourgogne, lesquels choisiriez-vous ?
- sous-corpus « définition » : qui permet de verbaliser des traits définitoires à partir desquels les répondants conceptualisent l'objet Crémant de Bourgogne.

- Corpus *Questionnaire Crémant de Bourgogne*
  - Distribué par une agence de sondage : **2250** répondants.
- Corpus *Questionnaire espumante*
  - Distribué en ligne : **528** répondants.
- Critères sociogéographiques genre, âge et région équilibrés pour le Crémant.
- Catégories de consommation => accès aux questions ouvertes :
  - ✓ vin ;
  - ✓ vin pétillant, mousseux ;
  - ✓ champagne ;
  - ✓ spiritueux et alcools forts.

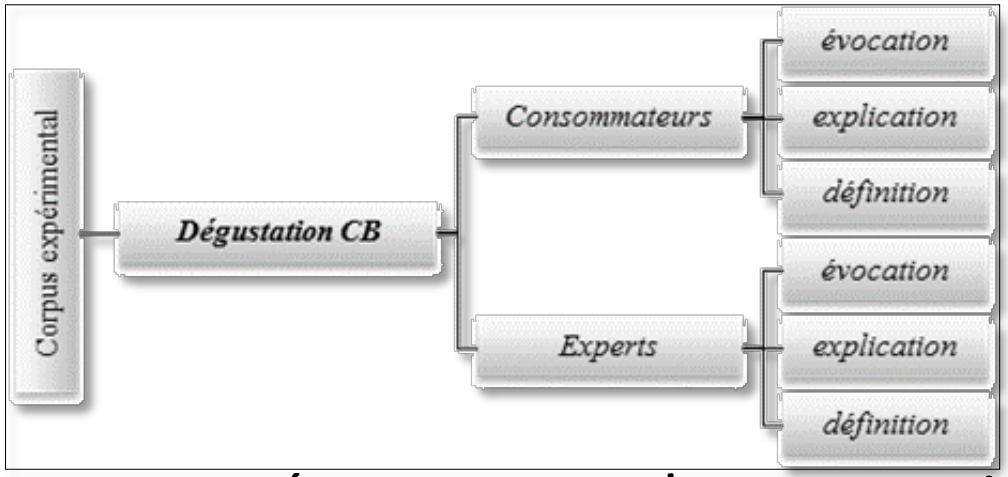
- Corpus *Questionnaire Crémant de Bourgogne* et *Questionnaire Espumantes*
- Données statistiques de base (AntConc) :

	Sous-corpus	Types	Hapax	Tokens
Questionnaire CB	• *évocation*	1 156	-	14 028
	• *explication*	1 330	-	19 996
	• *définition*	849	-	8 117
	TOTAL *Questionnaire CB*	**2 033**	**951**	**42 141**
	Sous-corpus	**Types**	**Hapax**	**Tokens**
Questionnaire espumantes	• *évocation*	864	-	3 649
	• *explication*	939	-	5 747
	• *définition*	514	-	2 160
	TOTAL *Questionnaire espumantes*	**1 560**	**835**	**11 556**

Le corpus **avec stimulus**

- Le corpus *Dégustation Crémant de Bourgogne*
  - Conçu en collaboration avec AgroSup Dijon et l'UPECB :
    - montée en gamme du Crémant de Bourgogne ;
    - évènement inédit intitulé « Les Éminents de Bourgogne » ;
    - une séance de dégustation originale réunissant ***consommateurs*** et ***experts***.
  - Il s'agit ici de faire répondre des **consommateurs** et des **experts** à des questions permettant de collecter un corpus à partir de la génération de paroles en situation de dégustation, où la mise en mots est couplée et déclenchée par le sensoriel.

- Mêmes trois questions => même trois sous-corpus de parole
  - 2 panels : expert et consommateurs



- Evènement centré sur une analyse sensorielle :
  - nécessité de collecter des données nombreuses ;
  - suivre un protocole de dégustation précis, basé sur des plans d'expériences afin de garantir la validité des résultats.

- Trois étapes :
  - deux étapes d'analyse sensorielle
  - une étape de constitution de données textuelles expérimentales (réponses au questionnaire)

1. Test hédonique à l'aveugle = j'aime ou je n'aime pas

2. Test bouteille et dégustation (conso) et Test par attributs (pro)

3. Mise en mots du Crémant de Bourgogne

- collecte de données à deux reprises : 2016 et 2017 à l'Auditorium de Dijon

- Corpus *dégustation Crémant de Bourgogne*
- Données statistiques de base (AntConc) :

	Sous-corpus	Types	Hapax	Tokens
Dégustation CB	***Consommateurs***	1 212	656	**8 286**
	• *évocation*	693	-	3 070
	• *explication*	735	-	3 948
	• *définition*	420	-	1 268
	Professionnels	975	-	**5 247**
	• *évocation*	509	-	1 916
	• *explication*	617	-	2 666
	• *définition*	252	-	665
	TOTAL *Dégustation CB*	**1 684**	**915**	**13 533**

6. Perspectives

- Corpus : notion incontournable en SDL, en particulier dans les approches se réclamant comme « cognitives » du fait du primat de l'usage…

- …mais : notion à interroger et à sortir d'approches strictement « mécanistes » …

- … ce qui peut passer, entre autres, par la fréquentation de corpus complexes et/ou non-conventionnels

Merci pour votre attention !

Laurent Gautier

Université Bourgogne Franche-Comté (EA4182)

laurent.gautier@ubfc.fr