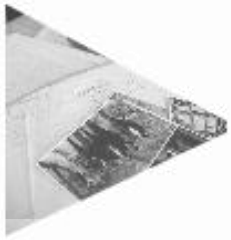

# Discours numériques : Quels enjeux pour la recherche en linguistique (appliquée)

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

# Structure

1. Eléments de contexte
2. Des corpus aux *digital data* en passant par les corpus numériques
3. Un défi pour la linguistique (appliquée) : quelques exemples
4. Perspectives

# 1. Eléments de contexte

**Tournant / révolution / transition numérique**

- Révolution des espaces de stockage et des temps de traitement des données => explosion des masses de données disponibles
- Appropriation du web 2.0 => nouvelles formes de communication avec apparition de nouveaux genres discursifs (au sens bakhtinien classique)
- Sous couvert de multimodalité => omniprésence des données langagières

# Exemple

# Les twitts, par-delà l'image / la vidéo / les liens

CCVF a retweeté

**Vin & Société** @vinetsociete · 17 oct.

Remise des prix #Œnovidéo : le programme court #1Minute1Vignoble reçoit le prix Adelphe/Anev ! cc @cnaoc bit.ly/2s7KMiO


Les Vins de Bordeaux, Vins de Provence, Vins de Bourgogne et 3 autres

1 11 19

# 2. Des corpus aux *digital data* en passant par les corpus numériques

**Les corpus et le quotidien du linguiste**

- Révolution des corpus pour l'approche de la réalité de la langue
- 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)

## Un changement de paradigme méthodologique du *corpus-based* au *corpus-driven*

- Sur corpus :
  - utilisation du corpus postérieure à la formulation des hypothèses
  - rôle essentiel de vérification/validation
- De corpus :
  - analyse du corpus antérieure à la formulation d'hypothèses
  - tout fait relevé doit être considéré comme pertinent
  - phénomènes absents aussi importants que phénomènes présents

# *Big* et *linked data* : un ’terrain de jeu’ pour linguistes

- Enjeux reposant sur une maîtrise et le développement d’outils linguistiques :
  - Extraction de l’information : traitement de la syntaxe (y compris des formes émergentes, par exemple liées au #), connaissances sémantiques, modélisation des savoirs encyclopédiques
  - Circulation de l’information : traitement de marqueurs de polyphonie, d’implicite, d’ironie, etc.
  - Représentation / visualisation de l’information : création de banques de données intelligentes, techniques d’annotation, d’apprentissage machine

# 3. Un défi pour la linguistique (appliquée) : quelques exemples

## L'analyse des discours évaluatifs pour l'extraction d'opinions

tripelle69
**Lyon, France**
145 51

Reviewed 6 days ago

bon hotel a paris

Google Translation

Enfin un hotel ou l'on se sent bien, belle chambre, propre, bien decorée, calme...seul bemol l'accueil qui laisse à desirer, manque de chaleur..; petit dejeuner fort agréable, peut être victime de leur succès les prix sont elevés et l'hotel souvent complet

- Essor des sites d'avis de consommateurs
- Traitement du lexique valencé (+, neutre, -) : création de dictionnaires appropriés + traitement de la 'tonalité' des textes
- Multiplication des domaines : tourisme, achats, finances

# L’analyse des réseaux sociaux pour l’extraction de l’information pertinente en


- Liens entre analyse linguistique et LSP : exemple #juratourisme
- Extraction de trois lexiques
  - Noms géographiques (en lien avec géolocalisation)
  - Lexique géographiques spécialisé
  - Lexique pertinent (*baignades*, *coins surpeuplés*) mais *a*

# L'analyse de blogs / forums pour l'extraction d'informations « terminologiques » en vue de la recommandation

Re: Domaine Morey-Coffinet

par **tpinault** » Mar 9 Fév 2010 13:18

Voilà donc le vin :

**Chassagne-Montrachet – Les Blanchots Dessus 2007**

Lez nez est boisé, trop, un peu de citron en arrière plan mais guère plus.
En bouche, c'est la banane et la vanille bourbon qui dominent. On sent tout de même de la tension dans ce vin et une finale intéressante.
Mais il n'est pas dans sa meilleure phase.
**Bien** –, à attendre impérativement ...

- Liens entre ontologie experte d'un domaine et « terminologie consos » pour recommander à partir des mots du consommateur
- Extraction et catégorisation du lexique de la dégustation avec triple dimension :
  - Sensoriel
  - Evaluative

# 4. Perspectives

« Le fait d'être appliquée donne à la linguistique une dimension pluridisciplinaire. » (Williams 2009 : 205-206)

- Interactions entre linguistique, informatique, sciences cognitives, (intelligence artificielle)
- Compilation de nouveaux types de données pouvant aussi servir à la modélisation de connaissances « théoriques »
- Collaboration (et financements…) avec les industriels de la langue

# Merci pour votre attention !

Pr Laurent Gautier
Université Bourgogne Franche-Comté (EA4182)
laurent.gautier@ubfc.fr