

# LeTALN et L'Europe

Christophe Chenon

# Le TALN, qu'est-ce que c'est ?

Traitement Automatique du Langage Naturel

( Natural Language Processing, NLP)

## ► Langue écrite

- Correction automatique, saisie prédictive, etc.
- Traduction automatique (TA, MT), résumés de texte, etc.
- Big data, informatique cognitive, données non structurées
  - classification, analyse

IA à base de règles >> statistiques >> réseaux neuronaux...

## ► Avec du traitement du signal, de l'image

- Dictée automatique, traduction de la parole
- Écriture manuscrite : adresses, chèques, saisie du chinois etc.

## ► Ergonomie (Interaction Homme-Machine, HCI)

- Agent conversationnel, commande vocale etc.
- Aide à la traduction (Traduction Assistée par Ordinateur, CAT)

# Diversité linguistique en Europe

## ► Les langues de l'Europe

- 24 langues officielles de l'Union
  - Entre autres : gaélique, maltais... et l'anglais !
  - Au moins une centaine de langues effectivement parlées

## ► Fonctionnement des institutions

- La langue de fonctionnement devient l'anglais
  - **2000** 50% des documents produits en anglais (1/3 en Français)
  - **2017** 75% en anglais ( moins de 10% en français)
- La BCE ne fonctionne qu'en anglais.
- Office Européen des Brevets : anglais français allemand.
  - 10 millions de brevets en 32 langues

## ► Intégration des populations et sentiment d'appartenance.

- Les députés du parlement parlent dans leurs langues
- 50% des Européens ne parlent qu'une seule langue

# La traduction en Europe

- ▶ **Direction Générale Traduction de la Commission**
  - ▶ Plus gros département
    - ▶ Dans le cadre de l'ISA (Interoperability Solutions for European Administration)
  - ▶ Traduction entre les 24 langues officielles
    - ▶ POETRY (flux) Euramis (mémoires) IATE (terminologie)
    - ▶ Un département par langue cible
- ▶ **Parlement européen**
  - ▶ 500 traducteurs
  - ▶ 30% du budget du parlement
- ▶ En tout pour l'UE : 1G € soit 2 euros par citoyen
- ▶ **MT@EC** (Traduction automatique de la CE, 2013)
  - ▶ Gratuit, financé par le CEF (Programme H2020)
  - ▶ Accessible aux administrations de l'UE ou locales

# La DG Connect (Ex DG XIII)

- ▶ Multilingual Digital Single Market (*Sommet de Riga 2015*)
  - ▶ Quelle politique de recherche dans le TAL pour promouvoir le **Marché unique numérique** (Single Digital Market )
  - ▶ Connecting Europe Facility,
    - ▶ Horizon 2020
- ▶  Multilingual Europe Technology Alliance
  - ▶ Financements de projets :
    - ▶ T4ME, META-NORD, CESAR, METANET4U
    - ▶ 60 laboratoires de recherche (34 pays)
  - ▶ Programme pour la recherche en TA
    - ▶ Surtout les langues peu dotées
    - ▶ Mesure de la qualité.
  - ▶ Effort partagé entre la CE et les états membres

# Conclusion

- ▶ La diversité linguistique est une spécificité européenne
  - ▶ La prééminence de l'anglais est manifeste
    - ▶ Fonctionnement de nombreux services
  - ▶ On ne peut pas fédérer les peuples par l'anglais
- ▶ L'activité de traduction est importante dans les institutions
  - ▶ La TA est indispensable
- ▶ La recherche est encouragée grâce par le CEF
  - ▶ Multilingual Digital Single Market
  - ▶ Programme Horizon 2020

[illegible]

# Les ressources linguistiques

- ▶ Les chercheurs manquent de ressources linguistiques
  - ▶ Sur 24 langues, 15 sont « peu dotées »
  - ▶ Problèmes d'éthique (utilisation des contenus privés)
  - ▶ Corpus parallèle Europarl
    - ▶ 50 millions de mots dans chaque langue
    - ▶ Très utilisé par les chercheurs.
- ▶ Bibliothèque numérique Européana
  - ▶ Commission Européenne (2008)
  - ▶ Référence les ressources numériques des bibliothèques nationales des 27 pays membres
  - ▶ 53 millions d'objets numériques référencés