

FOCUS 2

Syntaxe, sémantique lexicale et unités polylexicales

Emmanuel Schang (LLL), Denys Duchier (LIFO), Agata Savary (LIFAT), Caroline Pasquer (LIFAT), Cherifa Ben Khelil (LIFO)

Plan

Phénomènes

Données

Méthodes

Outils

Phénomènes

- Expressions polylexicales (EP) - séquences de mots qui ont une sémantique non-compositionnelle (le sens du tout ne peut pas être déduit des sens des composants de la manière régulière).

- We eat a big fish



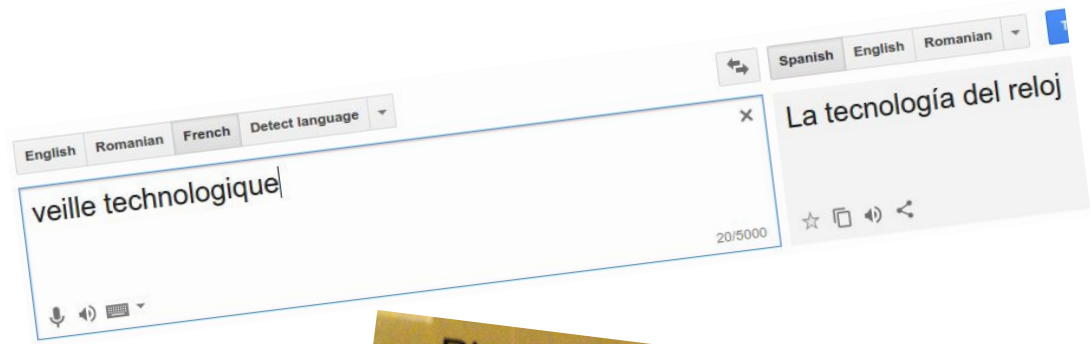
- We eat with a **big fish**



- Focus sur les expressions verbales:
 - à verbe support (**apporter** un **témoignage**),
 - expr. intrinsèquement réflexives (**s'emparer**),
 - idiomes verbaux (**mener** quelque chose **à bien**, **faire d'une pierre deux coup**)

Défi pour le TAL

- Les EP exigent des traitement dédiés pour éviter les grosses erreurs de traitement automatique (e.g. en traduction)



Données

- **Corpus PARSEME d'EP verbales (EPV):**
 - coordination du **réseaux** COST PARSEME (33 pays, 200 chercheurs)
 - terminologies et données **homogènes** en 20 langues, 6 millions de mots, 80,000 EPV annotées
- **Metagrammaires XMG**
 - métagrammaire des langues créoles (E. Schang)
 - métagrammaire du français (Crabbé et al.) étendue pour tenir compte des EP (A. Savary)

Méthodes

- **Coordination du réseaux COST PARSEME (par Blois et Marseille)**
 - effort collaboratif de 20 pays, homogénéisation des terminologies, guide d'annotation unifié, format et outils de traitement de corpus unifiés
 - Campagne d'évaluation PARSEME sur l'identification automatique d'EPV - 20 langues, 17 systèmes
- **Description par métagrammaires (XMG)**
 - description de haut niveau d'abstraction (indépendante du formalisme final)
 - codage lexical orienté objet
 - application aux langues créoles (E. Schang)
 - extension d'une métagrammaire du français (Crabbé, Duchier, et al.) aux EP (A. Savary)

Outils

- **varIDE** - système d'identification d'EPV:
 - parsing en dépendances, classification supervisée, focus sur les variantes syntaxiques des données vues
 - 5e dans le classement général de la shared task PARSEME
- **XMG + Tulipa**
 - formalisme d'une description méta-grammaticale
 - compilateur XMG vers LTAG (lexicalized tree-adjointing grammar)
 - parseur LTAG