Défis de l'enrichissement et du peuplement multilingue d'une ontologie à partir de corpus

Yuliya Korenchuk

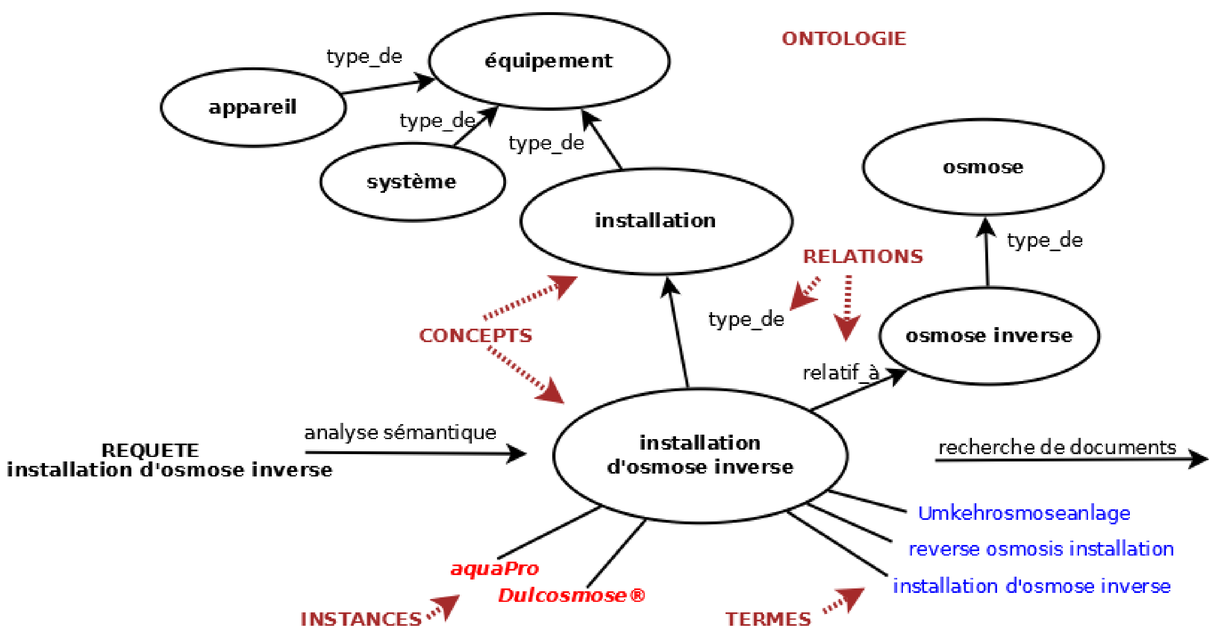
(LiLPa (Linguistique, Langues, Parole), EA 1339, Université de Strasbourg
yuliya.korenchuk@yahoo.fr

Exemple de recherche des informations multilingues

Requête : *installation d'osmose inverse*
Approche par mots-clés :
- Recherche dans la langue de la requête
- Uniquement les mots présents dans la requête

Approche sémantique :
- Prise en compte de la synonymie
- Recherche multilingue
- Prise en compte des instances

Ontologie pour la recherche d'informations multilingue



Ajout des informations dans une ontologie

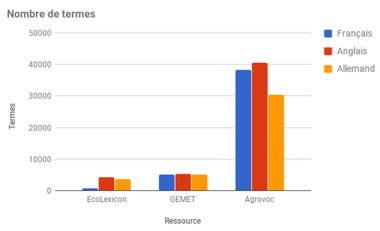
- **Enrichissement :** ajout des nouveaux concepts et relations [1]
- **Peuplement :** ajout des nouvelles instances [1]
- **Élargissement :** ajout des nouveaux termes (libellés) et instances (terminologie de Rebuz SAS)

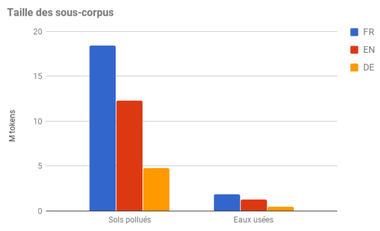
Défis d'enrichissement et d'élargissement multilingue

1. En fonction des **langues** et des **domaines** :
   - Volume variable de corpus
   - Disponibilité des ressources
   - Applicabilité des outils
2. **Évolution** des domaines :
   - Repérage des nouveaux termes et concepts
3. **Variation linguistique** dans le corpus :
   - Synonymie
   - Variantes d'orthographe
   - Aspect multilingue
4. Difficulté de **structuration** des connaissances

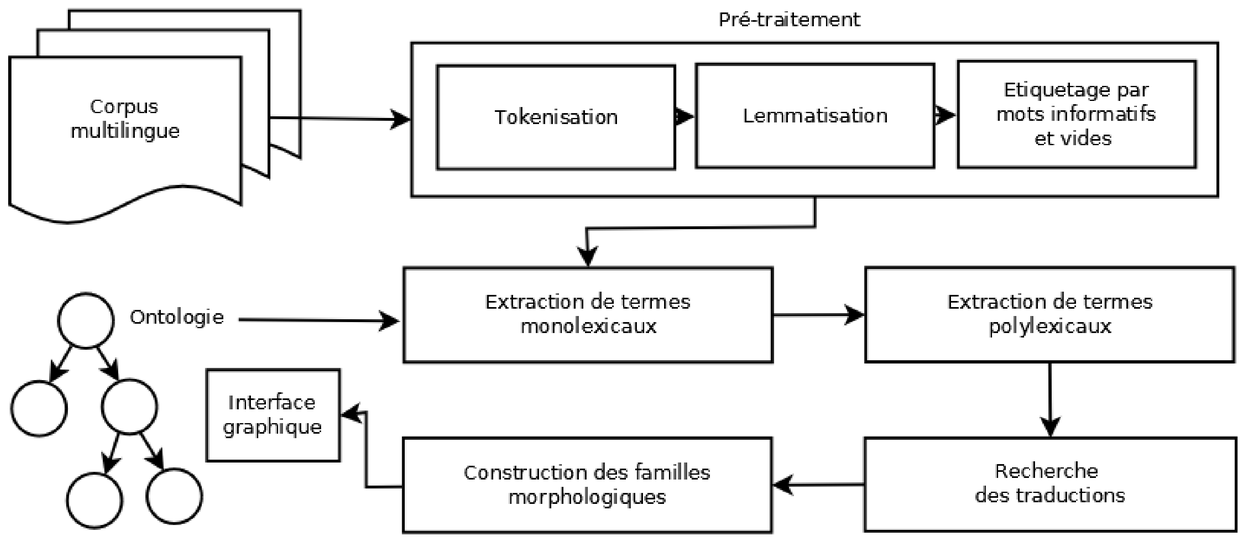
Ressources sémantiques et corpus

- 3 ressources sémantiques :
  - Ontologie EcoLexicon [2]
  - Thesaurus AGROVOC
  - Thesaurus GEMET
- Corpus des brevets PatTR [3]
- 2 sous-corpus :
  - Traitement des eaux usées
  - Traitement des sols pollués





Architecture



Solution : méthode basée sur les ressources endogènes

1. **N-grammes de caractères**
   - *osmose*, $n = 4$, : osmo, smos, mose, mose
2. **Étiquetage endogène de J. Vergne** [4]
   - *l_(n) invention_(n) être_(n)* ***particulièrement_(l)*** *efficace_(n) pour_(n)* ***assainir_(l)*** *le_(n)* ***biphényls_(l) polychloré_(l)*** *._(n)

- Ressources morphologiques
  - les n-grammes présents dans les libellés de l'ontologie
- Ressources morphosyntaxiques
  - les patrons endogènes de J. Vergne (InI, InII, etc.)
  - les n-grammes de la fin des mots (-tion, -ment, -ure, etc.)
  - les mots de schéma prédéfinis (de)
- Alignement bilingue des n-grammes de caractères (Anymalign [5])

Terme candidat FR	N-gramme	Score du n-gramme	Score du candidat
particulièrement	ière	0,19	0,01
	emen	0,06	
	icul	0,02	
assainir	sain	0,03	0,01
	ssai	0,03	
biphényls	phén	0,40	0,04
polychloré	chlo	0,20	0,03
	hlor	0,20	

TABLE: N-grammes permettant d'extraire les termes candidats

Terme candidat FR	Traductions candidates EN	DE
particulièrement	particularly	*fausse traduction*
assainir	*fausses traductions*	*fausse traduction*
biphényls	halobiphenyls polyhalogenatedbiphenyls	biphenyle biphenyl
polychloré	polychlorinate polychlorinate biphenyl	polychlorierte polychlorierten polychlorierter
biphényls polychloré	polychlorinate biphenyl	polychlorbiphenyl polychlorobiphenyl
assainir le biphényls polychloré	polychlorinate biphenyl	polychlorbiphenyl polychlorobiphenyl

TABLE: Traductions pour les termes candidats FR

Résultats : familles morphologiques multilingues endogènes

chlore@FR :
- *chloré@FR (chlorine-containing@EN) :*
  - *hydrocarbure chloré@FR*
  - *polychloré (polychlorinate@EN, polychlorierte@DE, polychlorierter@DE, polychlorierten@DE) :*
    - *biphényls polychloré@FR (polychlorinate biphenyl@EN, polychlorbiphenyl@DE* et *polychlorobiphenyl@DE)*
    - *biphényle polychloré@FR*
- *dichlorodiphényltrichloroéthane (dichlorodiphenyltrichloroethane@EN)*
- *chlorophosphate (chlorophosphate@EN)*, etc.

Conclusion

Méthode basée sur les ressources endogènes
- Robuste et nécessitant peu d'outils et ressources
- Multilingue, exploitant des corpus comparables
- Facilite le travail d'un expert humain sans le remplacer

Bibliographie

Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Anastasia Krithara, and Elias Zavitsanos.
Ontology Population and Enrichment : State of the Art.
In Georgios Paliouras, Constantine D. Spyropoulos, and George Tsatsaronis, editors, *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, pages 134–166. Springer Berlin Heidelberg, 2011.

Pamela Faber and Miriam Buendía Castro.
EcoLexicon.
Proceedings of the XVI EURALEX International Congress : The User in Focus, pages 601–608, 2014.

Katharina Wäschle and Stefan Riezler.
Structural and Topical Dimensions in Multi-Task Patent Translation.
In *Proceedings of EACL*, pages 818–828, Avignon, France, 2012.

Jacques Vergne.
Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée.
In *Actes de CIDE*, pages 155–168, 2005.

Adrien Lardilleux and Yves Lepage.
Anymalign : un outil d'alignement sous-phrastique libre pour les êtres humains.
In *Actes de TALN*, pages 24–26, 2009.

Remerciements : Nous remercions l'équipe LexiCon de l'Université de Grenade pour l'accès à la ressource EcoLexicon. Notre projet de recherche a bénéficié du financement CIFRE 2013/0744 accordé par l'ANRT (LiLPa/Rebuz SAS).
Je remercie mon employeur VERSUSMIND pour le financement de cette intervention.