ParCoGLiJe Corpus parallèle pour l'étude des grands classiques de la littérature de jeunesse Objectif du projet

Dejan Stosic, Saša Marjanović, Aleksandra Miletic

ParCoGLiJe

Corpus parallèle pour l'étude des grands classiques de la littérature de jeunesse

Dejan Stosic[1], Saša Marjanović[2], Aleksandra Miletic[1]

[1] CLLE, Université de Toulouse, CNRS, Toulouse, France | [2] Faculté de philologie de l'Université de Belgrade, Serbie

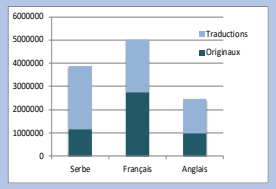
Journée d'étude CORLI
Université Paris Diderot - Paris 7
Traitements et standardisation des corpus multimodaux et web 2.0.
25 mai 2018

ParCoGLiJe: un dérivé de ParCoLab

ParCoLab est un corpus parallèle trilingue comportant des textes originaux et leurs traductions dans trois langues d'Europe: le français, le serbe et l'anglais. La ressource est développée et maintenue au sein du Laboratoire CLLE (UMR 5263, CNRS et Université Toulouse Jean Jaurès), en étroite collaboration avec la Faculté de philologie de l'Université de Belgrade (cf. Miletic et al. 2017).

Contenu

Le corpus ParCoLab comporte 11.359.761 mots provenant essentiellement de textes littéraires, mais aussi de textes juridiques, de sites web, de la presse, de retranscriptions de films, de divers textes spécialisés… La lemmatisation et l'annotation en POS et en fonctions syntaxiques sont en cours.



Accès et interrogation

☞ Consultable gratuitement en ligne à l'adresse: http://parcolab.univ-tlse2.fr/

☞Doté d'une interface d'interrogation permettant d'effectuer des requêtes simples ou complexes en combinant différents types de critères (expressions, lemmes, annotations morpho-syntaxiques et syntaxiques, etc.)

Hébergé par: Huma-Num — la TGIR des humanités numériques

Objectif du projet

Fournir un premier corpus parallèle français-anglais de littérature de jeunesse à tous ceux qui en font leur objet d'étude ou d'intérêt (littéraires, linguistes, traductologues, didacticiens, enseignants, grand public). Bien que la littérature de jeunesse soit un des domaines fondamentaux de la littérature et en dépit de sa place dans l'enseignement, la traduction en littérature pour la jeunesse reste un champ d'investigation très peu exploré (Lévêque 2014).

ParCoGLiJe

Huit grands classiques de littérature de jeunesse structurés au format XML (TEI-P5) alignées en français et en anglais au niveau des chapitres, des paragraphes et des phrases.

- **4 œuvres d'auteurs français**

Auteur	Titre	Taille FR	Taille EN	Formats
Daudet, A.	*Lettres de mon moulin*	46 592	47 706	xml, tmx, parcolab, bi-text
Dumas, A.	*Les trois mousquetaires*	213 791	228 900	xml, parcolab
De Ségur	*Mémoires d'un âne*	54 662	42 040	xml, parcolab
Verne, J.	*Vingt mille lieues sous les mers*	142 959	141 936	xml, parcolab

- **4 œuvres d'auteurs anglais**

Auteur	Titre	Taille FR	Taille EN	Formats
Hodgson Burnett, F.	*The Secret Garden*	76 940	80 558	xml, parcolab
Stevenson, R.L.	*Treasure Island*	69 827	68 996	xml, parcolab
Kipling, R.	*Jungle Book*	55 913	51 334	xml, parcolab
Dickens, Ch.	*Oliver Twist*	164 786	157 584	xml, parcolab

Taille du corpus: 1.644.524 mots

Un corpus à usages multiples

Bi-textes
- ✓ Grand public
- ✓ Littéraires comparatistes
- ✓ Traductologues

Lecture en L2 assistée
- ✓ Apprenants L2
- ✓ Grand public

Interrogation ParCoLab
- ✓ Linguistes
- ✓ Lexicographes
- ✓ Traducteurs
- ✓ Apprenants L2
- ✓ Enseignants FLE & ALE

XML standardisé
- ✓ Spécialistes TAL
- ✓ TAO

Diffusion du corpus

- ✓ **Disponible sur deux plateformes**

PARCOLAB & Ortolang

- ✓ **Diffusé en plusieurs formats (finalisation prévue pour septembre 2018)**
- ✓ **Diffusion de l'information (septembre 2018)**

Actions en cours et à venir

- ✓ **Enrichissement** de la ressource par d'autres œuvres, plus récentes
- ✓ **Finalisation du formatage** afin de rendre tous les textes disponibles dans tous les formats cités
- ✓ **Annotations** morphosyntaxiques et syntaxiques

Références

Lévêque, M. (2014), « Panorama de la recherche en littérature de jeunesse en France, 2013-2014 ». *In* Scherer, L., Issler, L. (éds), *Kinder- und Jugendliteratur der Romania. Impulse für ein neues romanistisches Forschungsfeld*, Peter Lang.

Miletic A., Stosic D., Marjanović S. (2017). « ParCoLab: A Parallel Corpus for Serbian, French and English ». In: Ekštein K., Matoušek V. (eds) *Text, Speech, and Dialogue. TSD 2017*. Lecture Notes in Computer Science book series, vol 10415. Springer, Cham, p. 156-164.

Remerciements et contacts

Nous remercions le consortium CORLI de la TGIR HumaNum, l'Equipex ORTOLANG, la plate-forme COCOON et le service des documents sonore de la BnF pour le financement qui a permis la réalisation de la ressource.

Contacts:
D. Stosic: dejan.stosic@univ-tlse2.fr
S. Marjanović: sasa.marjanovic@fil.bg.ac.rs
A. Miletic: aleksandramiletic1207@gmail.com