Ortolang
Open Resources and Tools for LANGuage
Outils et Ressources pour un Traitement Optimisé de la LANGue

ORTOLANG

Etienne Petitjean - ATILF - CNRS / Université de Lorraine
Cyril Pestel - ATILF - CNRS / Université de Lorraine
Christophe Parisse – MODYCO – CNRS/Université Paris Nanterre

Historique

- Projet monté en 2011
- En réponse à l'appel d'offre Equipex dans le cadre des PIA2
- Un équipement d'excellence de mutualisation de ressources et d'outils sur la langue et son traitement informatique
- Budget total de 2,6M € sur 7 ans
- Dès le départ vision nationale et internationale
- Volonté marquée d'axer le projet sur les données ouvertes

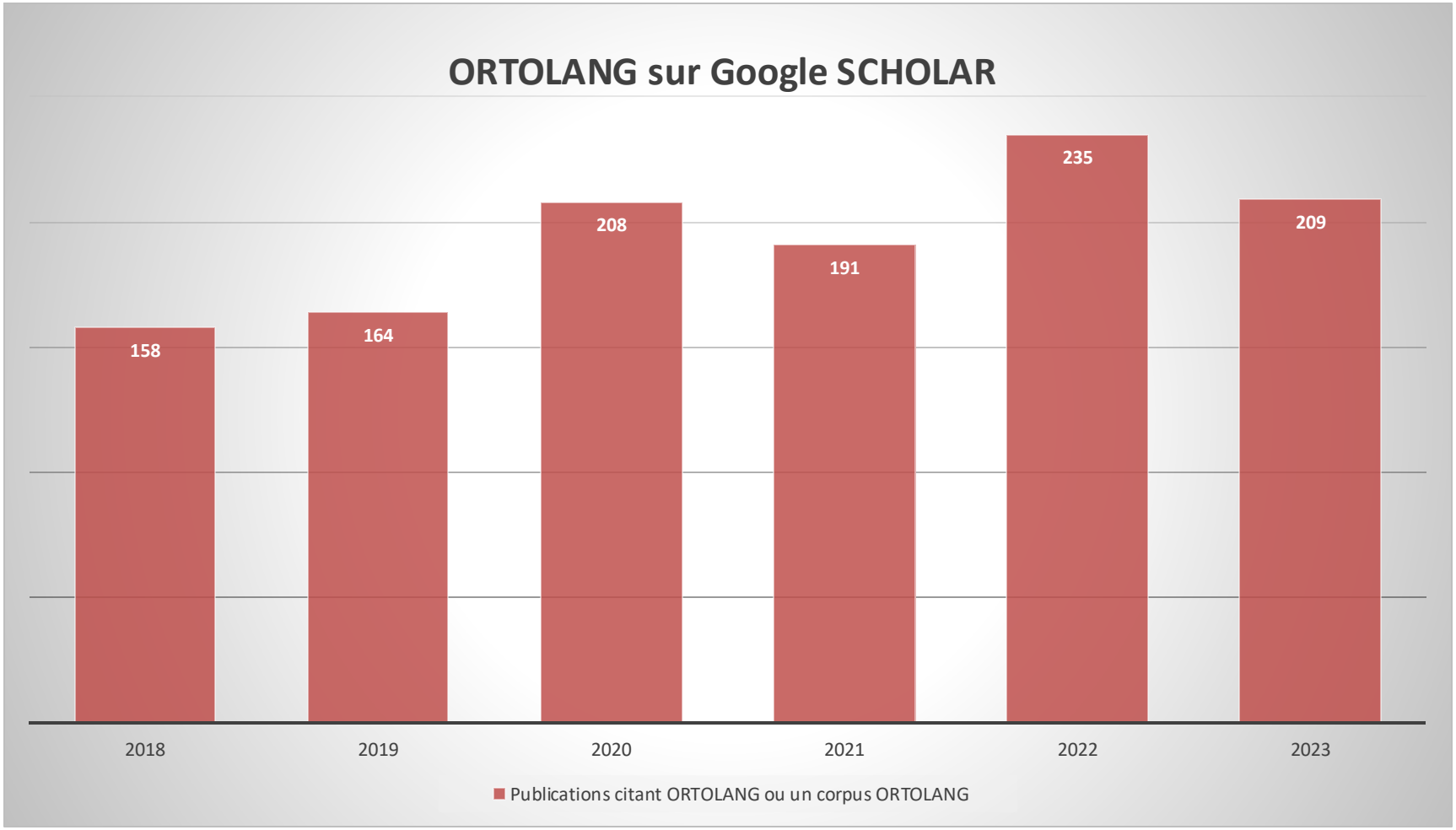
un consortium réunissant des compétences complémentaires

Objectifs

- Mutualisation de ressources et d'outils pour des travaux de recherche :
  - les ressources et les outils restent la propriété des laboratoires.
- Sécurité des données :
  - Authentification et droits d'accès
- Valorisation des données de la recherche :
  - FAIR
- Science Ouverte :
  - Encourager au maximum les données ouvertes
  - **« Aussi ouvert que possible, aussi fermé que nécessaire »**

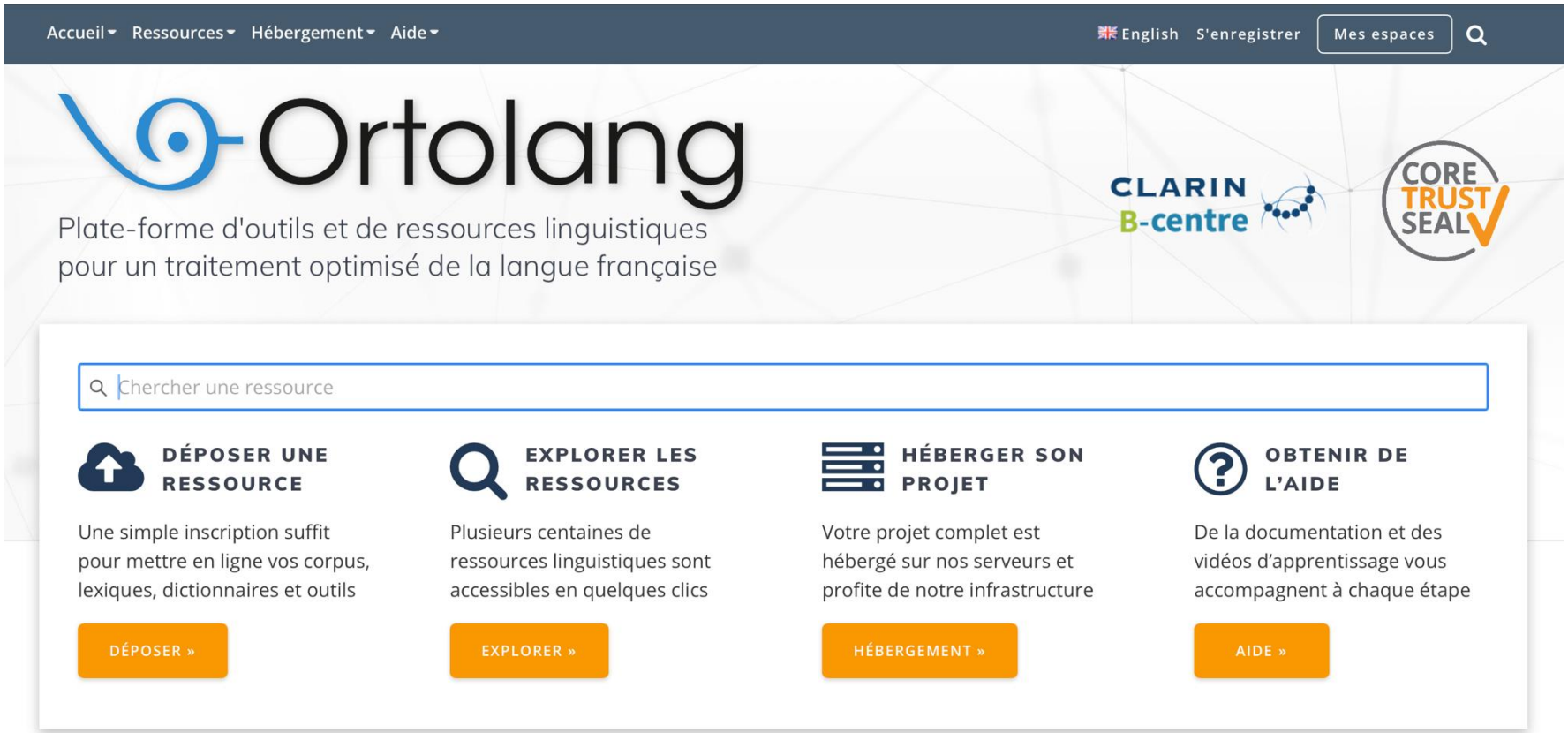
Statistiques

Visibilité (Google Scholar)

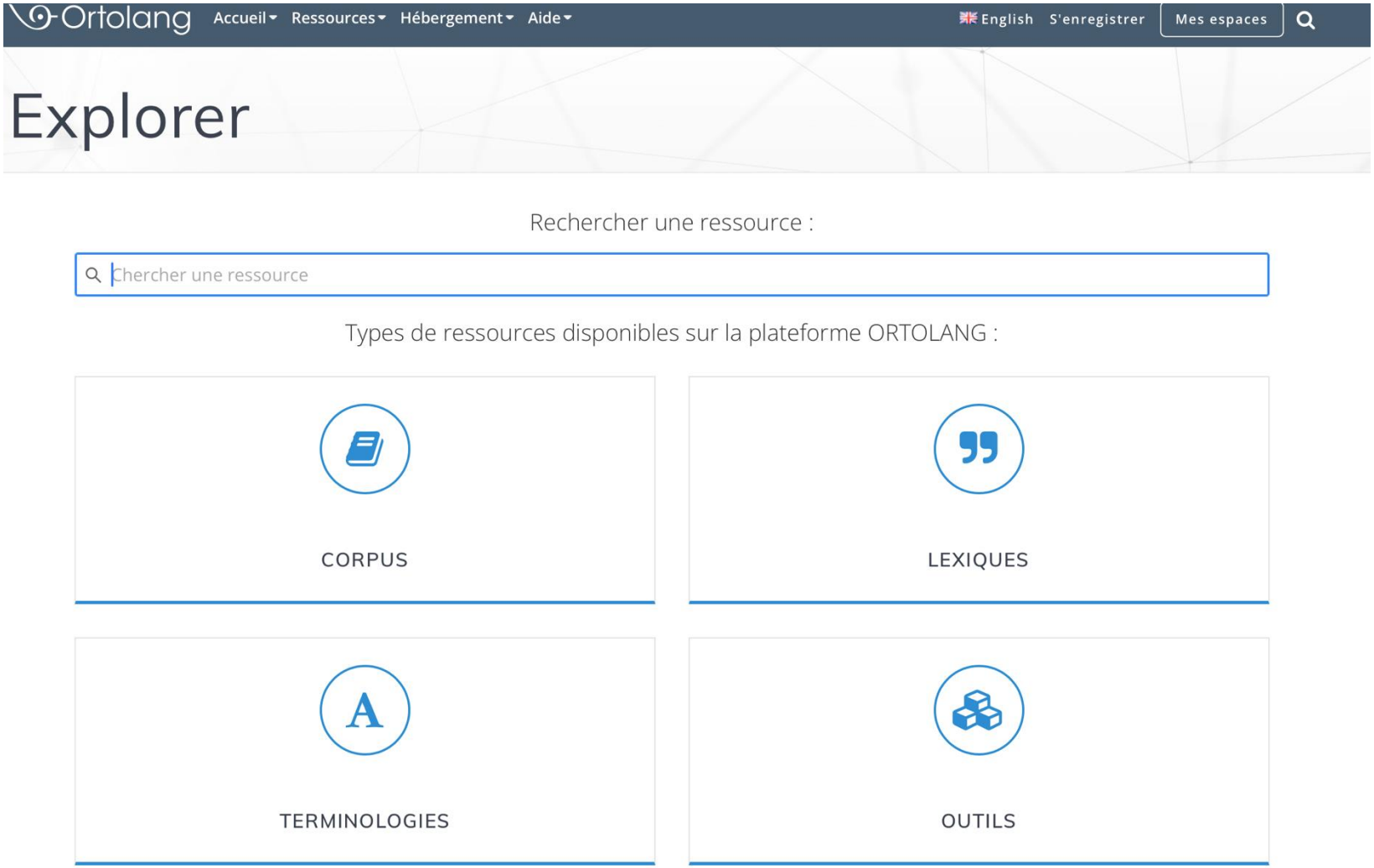


Utilisation d'ORTOLANG

- Accès par l'adresse **ortolang.fr**



Explorer


Ortolang
Accueil
Ressources
Hébergement
Aide
English
S'enregistrer
Mes espaces
Explorer
Rechercher une ressource :
Chercher une ressource
Types de ressources disponibles sur la plateforme ORTOLANG :
CORPUS
LEXIQUES
TERMINOLOGIES
OUTILS

Les corpus

ORTOLANG Catalogue Aide Langue Se connecter S'inscr

Rechercher un corpus

Corpus écrits (77 ressources)

Co		Di	co	Co			
Corpus lexical du farsi (thèse Jaferian 2024)	Climate discourses	Dictionnaire Étymologique Roman (DÉRom)	corpus_textuel_these_soni	Corpus Philéduc	EFG_WikiCorpus -- discussions en coulisse de Wikipedia (anglais, français, allemand)	MoNoPoli - Mots construits sur Noms propres de personnalités Politiques	BEL-RL-fr
2 oct. 2024	*9 sept. 2024*	*22 août 2024*	*24 juil. 2024*	*23 juil. 2024*	*21 juil. 2024*	*4 juil. 2024*	*19 juin 2024*

Corpus oraux (193 ressources)

				ORFÉO	TCOF	TA	Co
The CLES corpus of spontaneous L2 English	Corpus Finite Story: structure informationnellle et organisation du discours dans différentes langues et en L2	Interactions didactiques dans les disciplines scolaires	Corpus GEACC - Lab-FEACC	CEFC	TCOF : Traitement de Corpus Oraux en Français	TAM-MOUV	Corpus ANCOR Centre Version TEI
30 juil. 2024	*[illegible] 2024*	*27 mai 2024*	*1 mai 2024*	*10 avr. 2024*	*9 avr. 2024*	*25 mars 2024*	*13 mars 2024*

Corpus multimodaux (53 ressources)

		Di				Pr	
VIntAGe - Videos to study Interaction in AGEing	CLeLfPC	Dinlang_Doing_Attending	40 brèves	Mediapi-RGB	CIENSFO (Corpus d'Interrogatives Enchâssées Non-Standards du Français Oral)	Propicto	SMYLE
16 oct. 2024	*31 juil. 2024*	*1 juil. 2024*	*28 mai 2024*	*18 avr. 2024*	*10 avr. 2024*	*9 avr. 2024*	*15 août 2023*

Recherche pour « Modyco »

MoDyCo | Modèles, Dynamiques, Corpus - UMR 7114

Nanterre, FR

http://www.modyco.fr/fr/

Liste des productions du laboratoire

teicorpo

19 juil. 2016

PFC - Phonologie du Français Contemporain

2 mars 2017

NARRACOM

3 avr. 2018

TREEBANK RHAPSODIE

26 juin 2018

CEFC-GOLD

24 oct. 2019

VerNom

27 mai 2020

Shrug

24 juin 2020

Embodying Language Complexity: Co-speech gestures between 3 and 4

13 avr. 2021

Signes en famille

18 oct. 2021

Dictionnaire Informatisé des Mots d'Affect

19 janv. 2023

TREMoLo_Tweets_Corpus

21 mars 2023

DinLang - GESPIN 2023

24 mars 2023

Colaje

24 oct. 2023

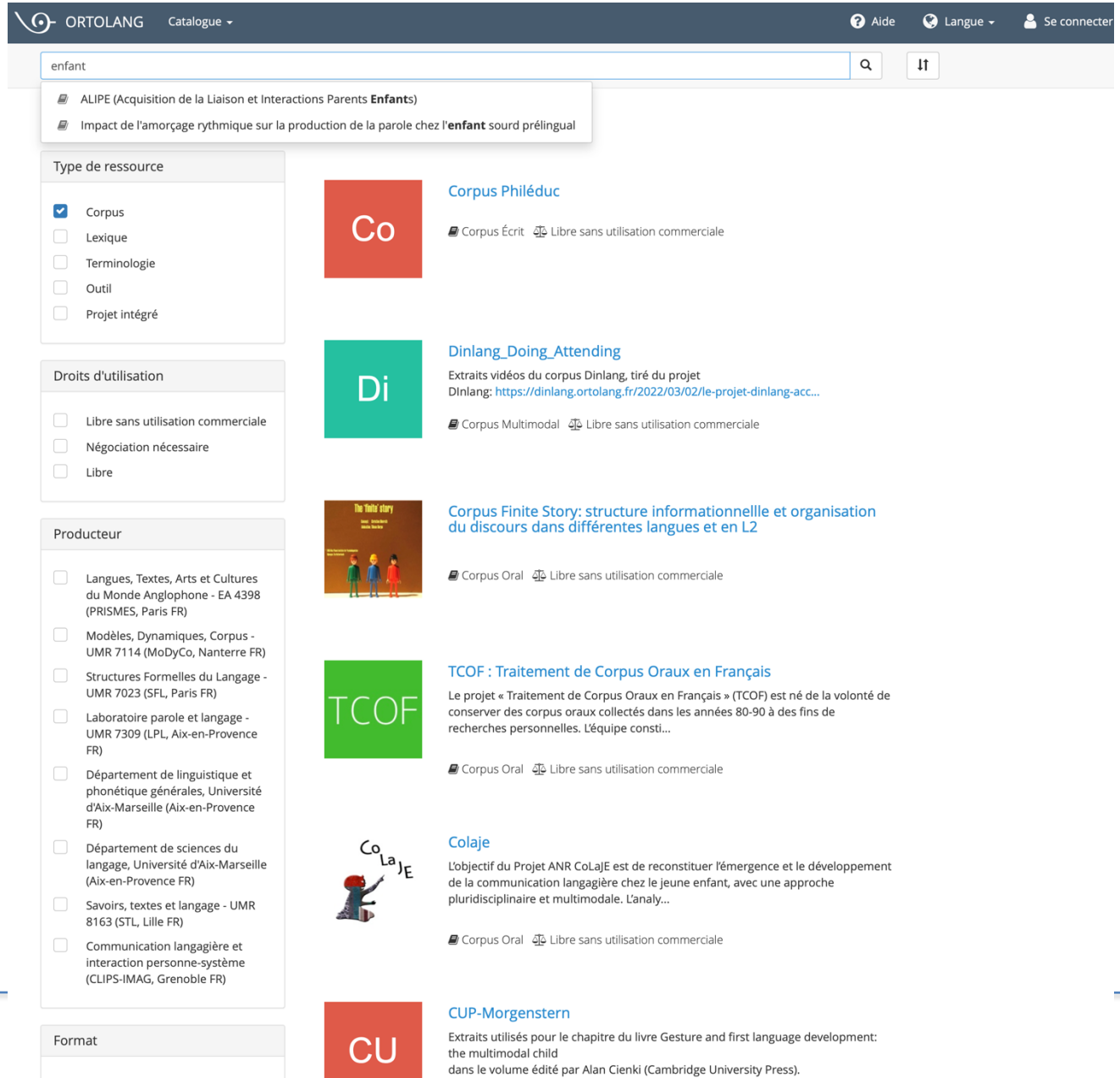
Dinlang_Doing_Attending

1 juil. 2024

VIntAGe - Videos to study Interaction in AGEing

16 oct. 2024

Recherche pour « Enfant »



Déposer

- S’identifier
  - Créer un utilisateur
  - Utiliser les identifiants de Renater
    - Identification comme Enseignement et Recherche (ESR)
- Créer et remplir un espace
- Publier cet espace
  - Un dépôt pérenne et un versionnage
- Revenir modifier et republier cet espace
  - Les anciennes versions restent accessibles

S’identifier

- 4 niveaux de protection des données
- Libre (pas d’identification) – données totalement ouvertes
- Utilisateurs identifiés → doivent avoir un compte sur ORTOLANG
  - Ouvert à toute personne
- Utilisateurs ESR → doivent appartenir à une institution d’enseignement supérieur ou de recherche (y compris doctorants)
- Privé → réservé aux données en cours de préparation ou aux données privées
  - L’utilisation de données privées doit être justifiée

Déposer et décrire des données

- Description complète dans https://www.ortolang.fr/fr/deposer/
- Créer un espace de travail
  - Le nommer (nom du corpus)
  - Décrire les métadonnées du corpus
  - Déposer les données
    - Glisser déposer, téléchargement (données isolées ou fichiers zip) et dépôt FTP pour les très grosses données
    - Pas de limite de format mais les données ouvertes sont recommandées
  - Donner les droits d'utilisation

Curation des données (modération)

- Une vérification des données est faite avant publication
  - Vérifier que les métadonnées sont correctes
    - Propriétaire et déposants
    - Laboratoires
    - Licence
    - Formats (formats ouverts demandés)
  - Se conformer à la charte d'ORTOLANG
    - Données ouvertes dans des formats connus et ouverts
  - Spécifier explicitement les limitations éventuelles
    - Données sous quarantaine, droits limités, condition d'utilisation

Accès aux données

- Données moissonnables par les moteurs de recherche comme Isidore, Clarin (VLO, Content search)
- Seules les pages principales sont accessibles par Google
- Téléchargement possible de tout un corpus, d'une partie de corpus, ou d'un fichier particulier.
- Tous les documents ont une adresse pérenne
  - Les fichiers peuvent être utilisés pour être accédés depuis un site web
  - https://hdl.handle.net/11403/cefc-orfeo/v1/oral/fleuron/accueil_etudiants_nancy2_partie2.orfeo
  - https://hdl.handle.net/11403/cefc-orfeo/v1/oral/fleuron/accueil_etudiants_nancy2_partie2.wav

Déposez vos données !!!

Évolutions récentes

- Certification CoreTrustSeal
- Certification Centre-B CLARIN
- Plateforme recherche.data.gouv.fr

CoreTrustSeal

- Certification internationale pour les entrepôts de données numériques
- « Self assessment »
- Formulaire de 16 questions:
  - Questions institutionnelles
  - Positionnement par rapport à une communauté
  - Qualité & sécurité des données
  - Qualité & sécurité des métadonnées
  - Documentation des procédures
  - Licences
  - Etc.

CoreTrustSeal (2)

- Plusieurs semaines de travail:
  - Rédaction en anglais (ATILF, INIST, Modyco)
  - Réflexion sur notre fonctionnement et nos procédures
  - Traduction de toute la documentation en anglais
  - Création d'un site multilingue de documentation

- Certification obtenue en aout 2023 pour 3 ans

CLARIN

- Infrastructure de recherche européenne sur les données langagières
- Participation de 24 pays
- Réseau de centres « nœuds »
  - B, C, K
- Notre objectif:
  - Devenir le premier Centre B français (22 au total)
  - Renforcer la visibilité internationale d'ORTOLANG

CLARIN (2)

- Certification technique:
  - CoreTrustSeal
  - Intégrer la fédération d'identité EduGain
  - Gérer des identifiants pérennes (PIDs)
  - Etc.

- Certification obtenue en mars 2024

CLARIN B-centre